Preliminary results on the iterative convergence of a class of implicit schemes

Jean-Antoine Desideri

INRIA
CENTRE
SOPHIA ANTIPOLIS

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P. 105
78153 Le Chesnay Cedex
France
Tél. (3) 954 90 20

Rapports de Recherche

N° 490

PRELIMINARY RESULTS ON THE ITERATIVE CONVERGENCE OF A CLASS OF IMPLICIT SCHEMES

Jean-Antoine DESIDERI

Février 1986

PRELIMINARY RESULTS ON THE ITERATIVE CONVERGENCE OF A CLASS OF IMPLICIT SCHEMES

Jean-Antoine DESIDERI

INRIA
Route des Lucioles
Sophia Antipolis
06560 VALBONNE

ABSTRACT

The iterative convergence properties of the fully implicit method is explored in the case where the explicit "physical phase" is a second-order accurate approximation of the inviscid gas-dynamic equations with an adjustable degree of upwinding, and the implicit "mathematical phase" is simplified to first-order accuracy. The investigation begins with an analysis in a general context, of the anomalies in the convergence of a linear iteration when the amplification matrix is defective. This allows us first to examine the finite-difference simulation of linear model problems, and second to discuss the results of a series of numerical experiments on the 2-d Euler equations using a finite-element program. One concludes that for large timesteps, fast convergence is achieved by either the simplified implicit method provided the explicit discretization averages central and fully upwind differencing, or by the regular implicit method with fully upwind second-order explicit and implicit discretizations.

RESUME

On explore les propriétés de convergence itérative de la méthode d'Euler implicite dans le cas où la "phase physique" explicite est une approximation du second-ordre des équations de la dynamique des gaz parfaits avec un degré ajustable de décentrage, et où la "phase mathématique" implicite est simplifiée au premier ordre. L'étude débute par l'analyse dans un contexte général, des anomalies de convergence d'une itération linéaire lorsque la matrice d'amplification est défective. Ceci permet d'étudier tout d'abord la simulation en différences finies de problèmes modèles linéaires, puis d'analyser une série d'expériences numériques portant sur les équations d'Euler 2-D en utilisant un programme d'éléments finis. La conclusion principale est que pour des grands pas de temps, une convergence rapide est réalisée soit par la méthode implicite simplifiée à condition que la discrétisation intervenant dans la phase explicite soit une moyenne entre un schéma centré et un schéma complètement décentré, soit par un schéma implicite complètement décentré dans sa phase explicite et sa phase implicite.

CONTENTS

1. INTRODUCTION

The Euler equations express the conservation of mass, momentum and energy in an inviscid fluid flow. Although viscous effects are often important in aerodynamics, these equations that account for the compressibility of the gas still represent a suitable physical model for a wide class of problems of interest in industry. For this reason, their numerical solution has received considerable attention in the past decade (–see for example recent issues of the AIAA Journal).

Here, we are considering the iterative convergence properties of a class of implicit schemes for the solution of the **steady Euler equations**. Although we are considering problems where the solution is independent of time, the time-dependent equations are employed in order to construct an iteration whose **converged solution** is the object of the calculation. Possible future applications may be three-dimensional and/or include general coordinate transformations. However, for the sole purpose of presentation, only the form taken by these equations in the case of two dimensions and Cartesian coordinates is recalled here ; that is, the following **'conservation-law form'**:

$$w_t + f_x + g_y = 0 \tag{1.1}$$

where the unknown vector w contains the so-called "conservative variables",

$$w = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ E \end{bmatrix} \tag{1.2}$$

where ρ is the density, u and v are the x and y velocity components and E is the total energy (internal+kinetic) per unit volume. In addition,

$$f = \begin{bmatrix} \rho u \\ \rho u^2+p \\ \rho uv \\ u(E+p) \end{bmatrix}, \quad g = \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2+p \\ v(E+p) \end{bmatrix} \tag{1.3}$$

where p is the pressure. For a perfect gas,

$$p = (\gamma-1)[E-\tfrac{1}{2}\rho(u^2+v^2)] \tag{1.4}$$

where γ is the ratio of specific heats (γ=1.4 for air). The substitution of (1.4) in (1.3) reveals that the **'flux vectors'** f and g are homogeneous functions of degree one in the components of w. As a result, defining the **'Jacobian matrices'**,

$$A(w) = \frac{\partial f}{\partial w}, \quad B(w) = \frac{\partial g}{\partial w} \tag{1.5}$$

we have, by application of Euler's theorem for homogeneous functions:

$$f(w) = A(w)\, w, \quad g(w) = B(w)\, w \tag{1.6}$$

Expressions for $A(w)$ and $B(w)$, the derivation of their basic properties as well as their utilization in flow simulations can be found in numerous places in the literature (see for example, [1-3]). The "conservation-law form" is employed here in view of applying the **'shock-capturing technique'**. For more details on these notions, the reader is referred to [4-5].

We introduce a space differencing operator P (whose precise definition will be made later) such that Pw is a discrete approximation of the space derivatives in (1.1), i.e.

$$\begin{aligned}(Pw)_{jk} &= (f_x)_{jk} + (g_y)_{jk} + \cdots \\ &= (Aw)_x|_{jk} + (Bw)_y|_{jk} + \cdots\end{aligned} \tag{1.7}$$

where ... indicates a discretization error. It follows that a space-discretized time-continuous analog of (1.1) is given by:

$$w_t + Pw = 0 \tag{1.8}$$

We recall that the matrices $A(w)$ and $B(w)$ can always be diagonalized and that their eigenvalues are real [1]. This results in the hyperbolic nature of (1.1) that can therefore be marched forward in time. A difficulty arises however unless the flow is entirely supersonic, that is, that these eigenvalues are not all of the same sign. This implies that a correct space-differencing scheme should either use central differences (or the like), or, in the case of an upwind scheme, should require flux-splitting [6]. To see this, we temporarily consider the case of an analogous linear hyperbolic test equation obtained from (1.1) by letting the flux vectors f and g be linear in w (i.e. A and B are constant). In this case the space differencing scheme represented by the operator P is constant in time, and besides the consistency condition, it seems natural to require that the exact solution of (1.8) namely $w(t)=exp(-Pt)w(0)$, should be bounded in time. But for the model problem, if the rule of space discretization cited above is respected, the real parts of the eigenvalues of the operator P are all nonnegative (if nonzero), and the requirement is met.

We now turn to the time-discretization method, sometimes referred to as the **'solver'**. In this study, the choice of the **'Backward Euler scheme'** also

known as the **'fully implicit method'** is made. When applied to (1.1) it can be written in the following **'delta form'**;

$$M(w^{n+1}-w^n) = -\Delta t\,(Pw)^n \tag{1.9}$$

where Δt is the time-step and

$$M = I+\Delta t\,(Pw)_w \tag{1.10}$$

(for a derivation see for example [7]). To evaluate the stability of this method we consider again the linear hyperbolic model, for which P and M can be thought as matrices constant during the iteration and

$$M = I+\Delta tP \tag{1.11}$$

so that, an amplification matrix $C(\Delta t)$ can be defined by

$$w^{n+1} = C(\Delta t)w^n \tag{1.12}$$

and turns out to be:

$$\begin{aligned} C(\Delta t) &= I-\Delta t\,(M^{-1}P) \\ &= I-[I+(\Delta tP)^{-1}]^{-1} \end{aligned} \tag{1.13}$$

Thus if the eigenvalues of P are denoted by $\{\lambda_m\}$ $(m=1,2,\ldots,N)$ those of $C(\Delta t)$ are given by

$$\begin{aligned} g_m(\Delta t) &= 1-\frac{1}{1+\frac{1}{z_m}} \\ &= \frac{1}{z_m+1} \end{aligned} \tag{1.14}$$

where $z_m = \lambda_m \Delta t$. Since for all m,

$$Re(z_m) \geq 0 \tag{1.15}$$

as justified previously in examining the solution of (1.8), it follows that for all Δt

$$|g_m(\Delta t)| \leq 1 \tag{1.16}$$

which establishes that the method is **unconditionally stable for the associated linear hyperbolic problem.** Furthermore,

$$\lim_{\Delta t\to\infty} g_m(\Delta t) = 0 \tag{1.17}$$

Of course these ideal results, valid for a linear model, may not all extend to the nonlinear case under study. However, we anticipate that using the Euler implicit method results in a method that is perhaps not unconditionally stable but at least **not limited by the CFL condition** [4-5] which restricts the usual explicit schemes, and that becomes **more dissipative with larger time-steps** (see Appendix 2 in [3]) without modifying the steady-state solution that remains governed by the operator P only.

Note that for an infinite time-step, if one lets $F(w) = Pw$, (1.9) becomes

$$F_w(w^n).(w^{n+1}-w^n) = -F(w^n) \tag{1.18}$$

In the above equation, we recognize **Newton's method** whose convergence is **quadratic** [8]. (For a linear problem, the amplification matrix G is equal to the null matrix and the iteration converges in one step). This confirms that being able to use stably very large time-steps is a highly desirable feature for the solver.

We now examine the algorithmic standpoint. The application of the algorithm defined in (1.9) is performed in three steps:

(1) **'Physical phase'**: evaluation of the vector $b = -\Delta t (Pw)^n$;

(2) **'Mathematical phase'**: solution of the system $M\ \Delta w^n = b$, in which the unknown is the vector Δw^n;

(3) **'Update'**: $w^{n+1} = w^n + \Delta w^n$.

The implicit mathematical phase preconditions the system in a way that enhances the stability of the method, but has no effect on steady-state accuracy. The physical phase however which is completed solely after proper boundary conditions are enforced, defines alone the converged solution. Therefore we require that the operator P be at least **second-order accurate** in regions where the solution is smooth. This is achieved by either a **central scheme** or the-like, or a second-order **upwind scheme**. The second alternative has gained some popularity in recent papers [9], because it yields schemes having (or almost having) certain desirable monotonicity properties and thus producing more physically relevant solutions near discontinuities, that is, optimally, accurate and oscillation-free solutions. A third alternative is to combine (linearly) a central discretization with an upwind discretization. In any case the requirement of obtaining a second-order accurate space discretization is generally not very difficult to meet because the step is explicit; moreover, in many (but not all) simple schemes, no Jacobians need be calculated but only linear combinations of the flux vectors f and g computed at various nodes, that is, the operator P is

not itself evaluated, but only and directly the vector $(Pw)^n$.

In contrast, the mathematical phase is far more complicated to realize, and this for several reasons:

–It is imperative to evaluate the operator M itself. (Evaluation of the elements of a matrix.)

–The calculation of the constituent blocks of M requires the computation of the 4x4 Jacobians (in 2-d), that is 4 times more (known) functions are involved than in the flux vectors.

–The system $M\,\Delta w^n = b$ need be solved. This is always a computationally difficult task when the system is large, particularly when the mesh is not regular in structure and smooth, or when the discretization is very sophisticated and complex. yielding a **large and ill-conditioned matrix (stiffness).**

–The inversion process necessitates the storage of the constituent blocks of the matrix M. When this matrix is too large, the limit of the computer's storage capability is attained and there is no alternative to overwriting certain blocks and reevaluating them everytime they reappear in a subsequent calculation. Thus some of the Jacobians are evaluated more than once, and more work is required than would be expected solely by inspection of the mathematical equations.

For all of these reasons, authors have proposed simplified versions of the implicit method, in which the Euler equations are approximated only to first-order in the mathematical phase, while the physical phase remains the same [9-10]. This reduces the bandwidth of the matrix M and thus significantly lessens the amount of computation done in the mathematical phase. But in doing so, a slight inconsistency in the formulation is introduced, and it is more than legitimate to question the real efficiency of the algorithm as an iterative scheme. In fact, preliminary experiments in which the explicit phase was a fully-upwind second-order scheme revealed to be deceiving from this standpoint.

These considerations motivated the present work, in which we analyze the **iterative convergence properties of the simplified Euler implicit method.** This is done first in Section 3, where the method is applied to model convection problems and evaluated both from the matrix theory standpoint and numerically, and second, in Section 4, for the 2-d Euler equations where a series of numerical experiments are reported. But prior to this, a few properties of linear iterations when the amplification matrix cannot be diagonalized are first established.

2. ON THE EFFECTS OF A DEFECTIVE AMPLIFICATION MATRIX

In this section, we examine the case of a general linear non-homogeneous iteration,

$$w^{n+1} = G\, w^n + b \tag{2.1}$$

in which the unknown w is an N-vector, G is a given $N \times N$ amplification matrix and b is a given constant N-vector. in the particular case where **the matrix G cannot be diagonalized.**

By a suitable similarity transform, G can still be reduced to the so-called **'Jordan canonical form'** [11]:

$$J = X^{-1}\, G\, X \tag{2.2}$$

where X is the generalized eigenvector matrix, and J is block-diagonal:

$$J = BDiag(J_i) = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_s \end{bmatrix} \tag{2.3}$$

where s is the number of linearly independent (true) eigenvectors ($s < N$), and each block J_i has the following bidiagonal structure:

$$J_i = Bidiag(1,\lambda_i) = \begin{bmatrix} \lambda_i & & & & \\ 1 & \lambda_i & & & \\ & 1 & \lambda_i & & \\ & & \ddots & \ddots & \\ & & & 1 & \lambda_i \end{bmatrix} \tag{2.4}$$

The matrix J being triangular, it contains its eigenvalues in its main diagonal; these are the numbers $\{\lambda_i\}$ $(i=1,2,..,s)$ that also are the eigenvalues of the matrix G to which J is similar. To assure convergence of the iteration, it is assumed that the amplification matrix G satisfies the spectral radius condition [12]:

$$\rho(G) = \underset{(i=1,2,\ldots,s)}{Max}(|\lambda_i|) < 1 \tag{2.5}$$

In this case the matrix $I-G$ is invertible, and therefore the iteration (2.1) admits a fixed-point solution $w^\infty = (I-G)^{-1}b$ that satisfies

$$w^\infty = G\, w^\infty + b \tag{2.6}$$

Then defining the error vector e^n by

$$e^n = w^n - w^\infty \tag{2.7}$$

and subtracting (2.6) from (2.1) it follows that e^n satisfies the following homogeneous linear iteration

$$e^{n+1} = G\, e^n \tag{2.8}$$

which implies that

$$e^n = G^n\, e^0 \tag{2.9}$$

But,

$$G^n = X\, J^n\, X^{-1} \tag{2.10}$$

and in the basis of the generalized eigenvectors, the error vector becomes,

$$\varepsilon^n = X^{-1}\, e^n \tag{2.11}$$

so that combining (2.2) with (2.8)-(2.10) yields the expression:

$$\varepsilon^n = J^n\, \varepsilon^0 \tag{2.12}$$

In addition, as a consequence of the block-diagonal structure of the matrix J given by (2.3),

$$J^n = BDiag\,(J_i^n) \tag{2.13}$$

It is therefore apparent that the attenuation with increasing n of the error-vector components is governed by the powers of the individual blocks J_i taken separately. For this reason, in what follows, and without great loss of generality, we consider the case of only one block ($s=1$; $J=J_1$; $\lambda_1=\lambda$). Using (2.4) several times successively, we obtain:

$$J^2 = \begin{bmatrix} \lambda^2 & & & & & \\ 2\lambda & \lambda^2 & & & & \\ 1 & 2\lambda & \lambda^2 & & & \\ & \cdots & \cdots & \cdots & & \\ & & 1 & 2\lambda & \lambda^2 \end{bmatrix} \tag{2.14}$$

$$J^3 = \begin{bmatrix} \lambda^3 & & & & & \\ 3\lambda^2 & \lambda^3 & & & & \\ 3\lambda & 3\lambda^2 & \lambda^3 & & & \\ 1 & 3\lambda & 3\lambda^2 & \lambda^3 & & \\ & \cdots & \cdots & \cdots & \cdots & \\ & & 1 & 3\lambda & 3\lambda^2 & \lambda^3 \end{bmatrix} \tag{2.15}$$

More generally, and for $n<N$, the block J^n is a banded lower-triangular matrix having n nonzero subdiagonals below the main diagonal and whose first column-vector is

$$1st\ column-vector\ of\ J^n = \begin{bmatrix} \lambda^n \\ n\lambda^{n-1} \\ C_n^2\lambda^{n-2} \\ \vdots \\ C_n^j\lambda^{n-j} \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{2.16}$$

where $C_n^j = n!/[j!(n-j)!]$. Evidently, each nonzero entry of the matrix J^n is one of the monomials appearing in the expansion of $(1+\lambda)^n$.

The above vector is precisely the value of ε^n if $\varepsilon^0 = e^1 \equiv (1,0,0,\ldots,0)^T$, that is the first vector of the canonical basis. Since one of its components (the $n+1$st) is equal to 1, its sup-norm (or maximum component in absolute value) is greater or equal to 1, and this, over at least $N-1$ iterations. Therefore, in the first $N-1$ iterations the process cannot be expected to be dissipative, even if the spectral radius is equal to 0.

For $n \geq N$, the above vector is truncated to the first N components. In particular, its last component becomes $C_n^{N-1}\lambda^{n-N+1}$. This term, for fixed N, is equal to the value at n of a monomial of degree $N-1$ times λ^n. The other components involve monomials of lesser degrees. Therefore, in the most general case, the vector ε^n is a general linear combination of the column-vectors of the matrix J^n and any of its components, say ε_j^n, is of the form $P_j^{N-1}(n)\lambda^n$ where P_j^{N-1} is a polynomial of degree $N-1$ depending on j. Consequently, as $n\to\infty$, $\|\varepsilon^n\|_\infty \to 0$ only at the rate of $n^{N-1}\rho^n$. From this we conclude that relatively to the regular case for which the amplification matrix G can be diagonalized and the error tends to 0 at the same rate as ρ^n, the asymptotic convergence is slightly degraded by the additional factor n^{N-1}. Unfortunately, another form of degradation of the convergence, believed to be more severe than the first, is now going to be demonstrated. For this, let:

$$\tau_n = \|\varepsilon^n\|_\infty = \underset{(j=0,1,2,\ldots,N-1)}{Max}(\xi_j^n) \tag{2.17}$$

where

$$\xi_j^n = C_n^j\rho^{n-j} \tag{2.18}$$

in which again, $\rho=\rho(C)=|\lambda|<1$, is the spectral radius. It turns out that the asymptotic convergence rate is significant only after a relatively large number of iterations. To illustrate the phenomenon, we begin with a numerical experiment. The sequence $\{\tau_n\}$ was evaluated in the particular case where $\rho=½$, $N=20$ and $\varepsilon^0=(1,0,...,0)^T$. The result of this calculation is shown on Figure 1 on a semi-logarithmic plot, where the sequence $\{2^{-n}\}$ is also represented. There it appears that the sequence $\{\tau_n\}$ is firstly constant and equal to one over a few iterations, then it is monotone-increasing over a number of iterations equal to about 100, that is $2N$, and finally it decreases, producing a parabolic branch admitting no asymptote but instead, an asymptotic direction that is the straight line representing 2^{-n} on this plot. (This agrees with the asymptotic convergence rate previously determined.) Also note that when τ_n achieves its maximum, the value of ρ^n is yet close to about 10^{-10} indicating that the spectral radius in this experiment was chosen relatively small, a generally very favorable circumstance. To explain these observations, we return to the general case and verify that:

For N large, it takes the sequence $\{\tau_n\}$ a number of iterations equivalent to $N/(1-\rho)$ to become monotone-decreasing.

In this analysis, N is large and since $n>N$, n is a fortiori large also.

First examine the variation with j of ξ_j^n for fixed n. We have:

$$\frac{\xi_j^n}{\xi_{j-1}^n}=\frac{n-j+1}{j\rho}>1 \quad \text{iff} \quad j<\frac{n+1}{\rho+1} \tag{2.19}$$

This leads us to analyze two cases separately:

1st case: $N<n<(\rho+1)N-1$ (*N large*).

In view of (2.19), it appears that since the ratio $(n+1)/(\rho+1)$ is less than N, it is for a value j_0 of j close to that ratio that ξ_j^n achieves its maximum over j, which implies that: $\tau_n=\xi_{j_0}^n$. The statement in (2.19) also indicates that if n increases of 1, the maximum will be achieved at $j_1=j_0$, or possibly j_0+1. Therefore $\tau_{n+1}/\tau_n=\xi_{j_0}^{n+1}/\xi_{j_0}^n$, or possibly $\xi_{j_0+1}^{n+1}/\xi_{j_0}^n$. In both cases, using $j_0/n\approx 1/(\rho+1)$, one obtains that $\tau_{n+1}/\tau_n\approx\rho+1>1$, and therefore n is found insufficiently large for the sequence $\{\tau_\nu\}$ $(\nu\geq n)$ to be monotone-decreasing.

2nd case: $n\geq(\rho+1)N-1$ (*N large*).

Then (2.19) indicates that the sequence $\{\xi_j^n\}$ increases with j.

Therefore, for all $\nu\geq n$, $\tau_\nu=\xi_{N-1}^\nu=C_\nu^{N-1}\rho^{\nu-N+1}$.

Consequently, $\tau_{\nu+1}/\tau_\nu=(\nu+1)\rho/(\nu+1-N+1)=\rho/[1-(N-1)/(\nu+1)]$,

and this ratio is less than 1 for all $\nu\geq n$ iff $n>(N-2+\rho)/(1-\rho)\sim N/(1-\rho)$. ▪

In conclusion, we summarize this section as follows: if a linear iteration has a defective amplification matrix G, still satisfying the spectral radius condition, $\rho<1$, that insures convergence, then for a general initial guess, the asymptotic convergence will only be like $n^{N-1}\rho^n$, where N is the dimension of the larger Jordan block appearing in the reduction of the amplification matrix. Moreover, if the number N is large, indicating that a large number of eigenvectors are missing, the asymptotic convergence rate is meaningful only after a number of iterations of the order of $N/(1-\rho)$, a particularly severe degradation if N is very large or if ρ is close to unity, or both.

In the next section, we examine the application of a particular implicit finite-difference scheme to model convection problems. For certain values of a parameter controlling the degree of 'upwinding' in the space discretization, the amplification matrix cannot be diagonalized and the effects of this are evaluated.

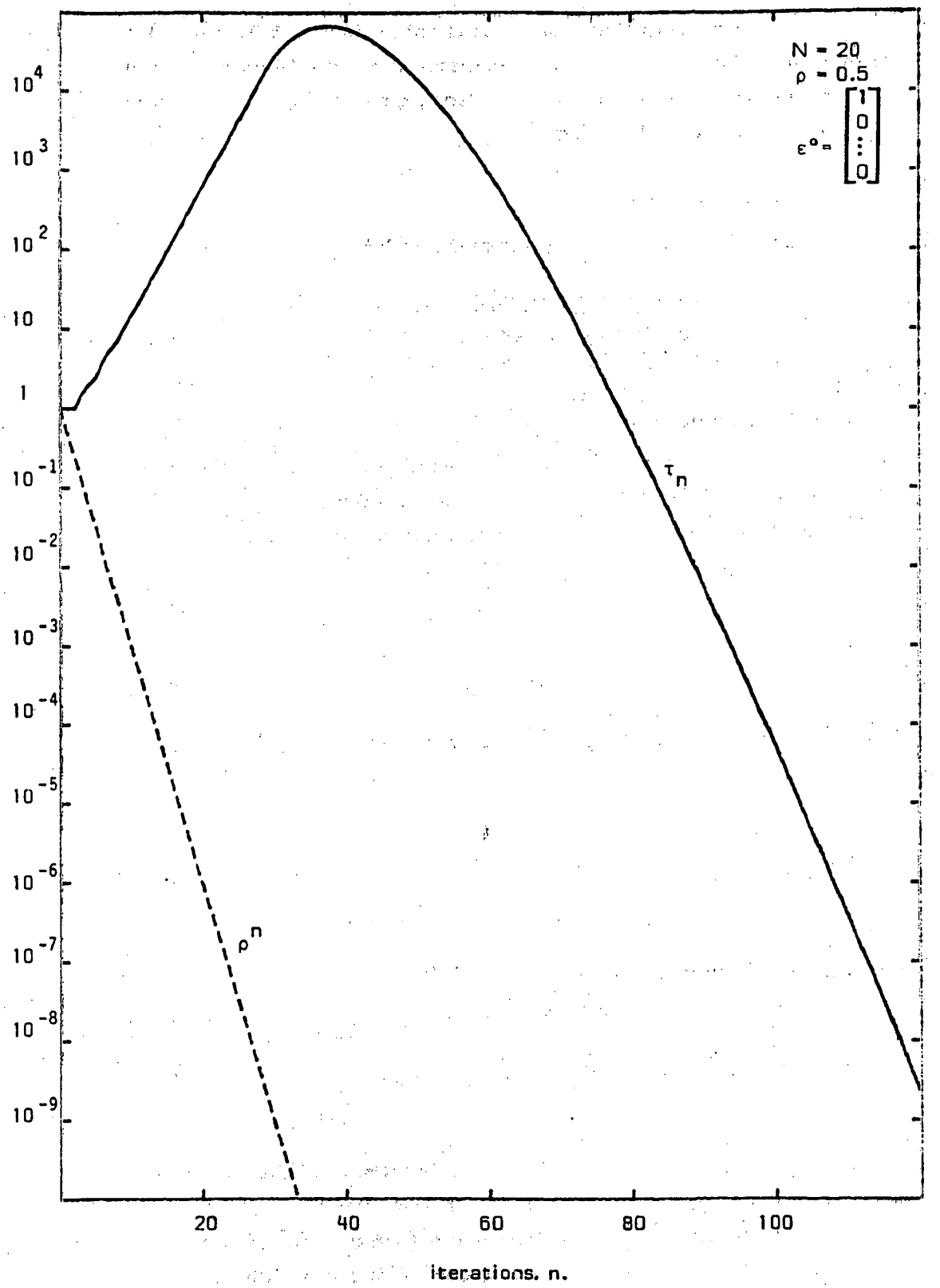

Figure 1. Variation of τ_n with n compared to ρ^n .

3. MODEL CONVECTION PROBLEMS

In this section, we investigate the properties of the simplified fully implicit method defined in Introduction applied to one-dimensional and two-dimensional scalar constant-coefficient linear convection problems, depending on the degree of upwinding appearing in the explicit phase.

3.1. One-dimensional model problem.

In this Subsection, we consider the quarter-plane problem:

$$\begin{cases} u_t + cu_x = 0 \ \ (c>0) \\ u(x,0) = u^0(x) \ \ (0 \leq x \leq L) \\ u(0,t) = u_0 \ \ (t>0) \end{cases} \tag{3.1}$$

which is often used for theoretical purposes.

The operator P mentioned in Introduction is a finite-difference discretization of the term cu_x. It is taken to be a linear combination of central differencing, δ^c, with the second-order backward-difference operator, δ^u. Thus:

$$P = (c/\Delta x)\delta_2 \tag{3.2}$$

with

$$\delta_2 = (1-\beta)\ \delta^c + \beta\ \delta^u \tag{3.3}$$

where δ^c is represented by the following (nearly skew-symmetric) tridiagonal matrix:

$$\delta^c = Trid(-\tfrac{1}{2},0,\tfrac{1}{2}) = \tfrac{1}{2}\begin{bmatrix} 0 & 1 & & & & \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & \cdots & \cdots & \cdots & \\ & & & -1 & 0 & 1 \\ & & & & -2 & 2 \end{bmatrix} \tag{3.4}$$

and δ^u the following lower-triangular tridiagonal matrix:

$$\delta^u = \tfrac{1}{2}\,LTrid(1,-4,3) = \tfrac{1}{2}\begin{bmatrix} 2 & & & & & \\ -4 & 3 & & & & \\ 1 & -4 & 3 & & & \\ & 1 & -4 & 3 & & \\ & & \cdots & \cdots & \cdots & \\ & & & 1 & -4 & 3 \end{bmatrix} \tag{3.5}$$

To arrive at (3.4)-(3.5) the boundary value u_0, which is known to have no effect on convergence was set equal to zero. Also, the first-order backward difference replaces the central difference at the last gridpoint (last row in (3.3)), and the second-order backward difference at the first gridpoint (first row in (3.4)).

In this way the parameter β controls the degree of 'upwinding' in the explicit phase. We now consider the mathematical phase. A difference operator simpler than P, that is with a reduced bandwidth, is the first-order accurate backward difference operator, which is represented by the following lower-triangular bidiagonal matrix:

$$\hat{P} = \left(\frac{c}{\Delta x}\right) \delta_1 \tag{3.6}$$

with

$$\delta_1 = Bidiag(-1,1) = \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \cdots & \cdots & \\ & & & -1 & 1 \end{bmatrix} \tag{3.7}$$

This operator is used in place of P to define the matrix M of the mathematical phase (see (1.11)). Consequently, the amplification matrix is no longer given by (1.13) but instead:

$$G(\Delta t) = I - \Delta t (I + \Delta t \hat{P})^{-1} P \tag{3.8}$$

As a result, when $\Delta t \to \infty$, the above matrix approaches a *nonzero* limit given by:

$$G_\infty = I - \delta_1^{-1} \delta_2 \tag{3.9}$$

This is a key formula throughout this report. It can easily be verified that the matrix associated with the operator δ_1^{-1} is a lower triangular matrix whose entries in the main diagonal and below are all equal to one:

$$\delta_1^{-1} = \begin{bmatrix} 1 & & & & \\ 1 & 1 & & & \\ 1 & 1 & 1 & & \\ \vdots & \vdots & \vdots & & \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix} \tag{3.10}$$

(Note that this is, not surprisingly, a first-order accurate discrete integration operator.)

Since both δ_1 and δ_2 are first-difference operators, that is, they both approximate the same operator, $\Delta x \partial / \partial x$, they are in this sense "close" to one another. Thus, when the discretization is simplified to first-order accuracy in the mathematical phase for algorithmic reasons, it is hoped that the amplification matrix G_∞ will still remain "close" to the null matrix, an enviable situation for fast convergence. The object of this report is essentially to precise the conditions under which this conjecture can be supported, from the *eigenvalue* standpoint, since the spectral radius must be small to achieve our goal, but also from the *eigenvector* standpoint, since in regard of the previous section, a defective

amplification matrix is not desirable.

We begin with the observation of a very simple result. Let V and V' be the following vectors:

$$V = C \begin{bmatrix} 1 \\ 2 \\ 3 \\ \vdots \\ N \end{bmatrix} \quad V' = C \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{3.11}$$

in which C is a constant. These vectors are discrete analogs of a linear function and its constant derivative respectively. The simple substitution of these definitions in the appropriate formulas gives:

$$\forall \beta, \ \delta_1 V = \delta^c V = \delta^u V = \delta_2 V = V' \tag{3.12}$$

Consequently,

$$\forall \beta, \ (\delta_2 - \delta_1) V = 0 \quad \text{and} \quad C_\infty V = 0 \tag{3.13}$$

Therefore:

Independently of β, $\lambda=0$ is an eigenvalue of the matrix C_∞. An associated eigenvector is the vector V of (3.11).

This was anticipated, since the operators δ_1 and δ_2 only differ by second-order terms, regardless the value of the parameter β, and it is only natural that the matrix C_∞ exactly anihilates a linear distribution of data. To make this more precise, note that interpreting δ_1 and δ_2 as point operators, then away from the boundaries, for some smooth function u and independently of β:

$$\delta_2 u \doteq \delta_1 u + \frac{\Delta x^2}{2} u_{xx} + O(\Delta x^3) \tag{3.14}$$

and substituting this in (3.9) gives:

$$C_\infty u = -\frac{\Delta x^2}{2} \delta_1^{-1} u_{xx} + O(\Delta x^3) \tag{3.15}$$

But, in view of (3.10), for any smooth function v:

$$\Delta x \ \delta_1^{-1} v = \int_0^x v dx + O(\Delta x) \tag{3.16}$$

and combining the last two equations yields:

$$C_\infty u = -\frac{\Delta x}{2} [u_x - u_x(0)] + O(\Delta x^2) \tag{3.17}$$

Therefore, the operator $-C_\infty$ may be interpreted as half a first-difference

operator. This interpretation will be used later.

We now proceed in the determination of the remaining eigenmodes, and for this several cases corresponding to different values of β are examined.

3.1.1. Central Differencing (β=0).

Here, $G_\infty = G_\infty^c$ where:

$$G_\infty^c = I - \delta_1^{-1}\delta^c = \tfrac{1}{2}\begin{bmatrix} 2 & -1 & & & & \\ 1 & 1 & -1 & & & \\ 1 & 0 & 1 & -1 & & \\ \vdots & \vdots & & \cdots & \cdots & \\ 1 & 0 & \cdots & 0 & 1 & -1 \\ 1 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \tag{3.18}$$

The characteristic polynomial of this $N \times N$ matrix is developed in Appendix A. There, it is found to be:

$$P_N^c(\lambda) = -\lambda(\tfrac{1}{2}-\lambda)^{N-1} \tag{3.19}$$

Consequently:

Besides the eigenvalue λ=0 which is simple, the matrix G_∞^c has only one eigenvalue, λ=½ of multiplicity $N-1$. (The spectral radius is thus ρ^c=½.)

To determine the associated eigenvector, we put $U=\{U_j\}$ into the relation $(G_\infty^c - \tfrac{1}{2}I_N)U=0$ and find:

$$\frac{U_1-U_2}{2} = \cdots = \frac{U_1-U_j}{2} = \cdots = \frac{U_1-U_N}{2} = \frac{U_1+U_{N-1}-2U_N}{2} = 0 \tag{3.20}$$

This gives U_j = a constant. Thus we find only one eigenvector associated with the multiple eigenvalue λ=½, that is $U = V$ of (3.11). Therefore:

The matrix G_∞^c is defective, $N-2$ eigenvectors are missing, the largest Jordan block being of order N-2.

Since N is large, we can combine the last two results with those of the previous section, and conclude:

For central differencing (β=0), and general initial guess, the iteration

$$U^{n+1} = G_\infty^c\, U^n + b \tag{3.21}$$

in which U^n is the n-th iterate (an N-vector), b is a given N-vector, and N is considered large, is non-dissipative over a number of iterations of the order of $2N$, and then enters the final phase of convergence which is asymptotically like $n^{N-2}/2^n$.

We recall that by *'non-dissipative iteration'* we mean that at least one component of the error-vector expressed in the basis of the generalized eigenvectors does not decrease monotonically when n increases.

One may question whether the situation could be improved by a more accurate (that is second-order) boundary differencing scheme. In Appendix B, it is shown that if the last row of the central-differencing scheme, δ^c, is replaced by the last row of the fully-upwind second-order scheme, δ^u, then the iterative convergence is not significantly improved, although the bandwidth of δ_2 (but not C_∞) is enhanced.

Lastly, when raising the matrix $(2\,C_\infty^c)$ to the power n, the coefficients of the binomial $(1-x)^n$ appear near the main diagonal with the same alternation of signs, and also near the first column (there, with the exception of the first one of them, that is, 1). This fact is clearly understandable from (3.17). Hence if we define the "highest-frequency mode" as the following vector W

$$W = (1,-1,1,-1,\ldots,(-1)^{N-1})^T \tag{3.22}$$

so that $\| W \|_\infty = 1$, we have

$$\forall\ n \geq 2,\quad \| (G_\infty^c)^n \|_\infty = \| (G_\infty^c)^n\ W \|_\infty = \frac{(2^n-1)+2^n}{2^n} = 2 - 2^{-n} \tag{3.23}$$

2.0.1. Fully-upwind second-order differencing (β=1).

Here, $G_\infty = G_\infty^u$ where:

$$C_\infty^u = I-\delta_1^{-1}\delta^u = \tfrac{1}{2}\begin{bmatrix} 0 & & & & & \\ 2 & -1 & & & & \\ 1 & 1 & -1 & & & \\ \vdots & \vdots & & \ddots & & \\ 1 & 0 & \cdots & 1 & -1 & \\ 1 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \tag{3.24}$$

Note that the first row of this matrix is made of zeros. This is because the accuracy of the backward-difference was dropped to first order at the first gridpoint in (3.5), that is, $\delta^u u_1 = \delta_1 u_1$ and the operator $I-\delta_1^{-1}$ anihilates exactly this vector.

In view of (3.24), the matrix C_∞^u is lower-triangular and its eigenvalues appear in the main diagonal. It follows that:

Besides the eigenvalue $\lambda=0$ *which is simple, the matrix* G^u_∞ *has only one eigenvalue,* $\lambda=-\frac{1}{2}$ *of multiplicity* $N-1$. *(The spectral radius is thus* $\rho^u=\frac{1}{2}$.*)*

To determine the associated eigenvector, we put $U=\{U_j\}$ into the relation $(G^u_\infty+\frac{1}{2}I_N)U=0$ and find:

$$U_1 = \frac{U_1+U_2}{2} = \cdots = \frac{U_1+U_j}{2} = \cdots = \frac{U_1+U_{N-1}}{2} = 0 \qquad (3.25)$$

This gives $U_j = 0$ for $j = 1,2,\ldots,N-1$. Thus we find only one eigenvector associated with the multiple eigenvalue $\lambda=-\frac{1}{2}$, that is $U = e^N \equiv (0,0,\ldots,0,1)^T$. Hence:

The matrix G^u_∞ *is defective,* $N-2$ *eigenvectors are missing, the largest Jordan block being of order N-2.*

Since N is large, we can combine the last two results with those of the previous section, and conclude:

For fully-upwind second-order differencing ($\beta=1$), *and general initial guess, the iteration*

$$U^{n+1} = G^u_\infty\, U^n + b \qquad (3.26)$$

in which U^n *is the n-th iterate (an N-vector), b is a given N-vector, and N is considered large, is non-dissipative over a number of iterations of the order of* $2N$, *and then enters the final phase of convergence which is asymptotically like* $n^{N-2}/2^n$.

Also, when raising the matrix $(2\,G^u_\infty)$ to the power n, the coefficients of the binomial $(1-x)^n$ appear this time only near the main diagonal and with the same alternation of signs. This is again clearly understandable from (3.17). Hence it is again the highest-frequency mode, $U = W$ of (3.22), that maximizes the norm $\| (G^u_\infty)^n\, U \|_\infty$ and:

$$\forall\, n \geq 2, \quad \| (G^u_\infty)^n \|_\infty = \| (G^u_\infty)^n\, W \|_\infty = \frac{2^n}{2^n} = 1 \qquad (3.27)$$

2.0.2. Schemes linearly combining the central with the fully-upwind differencing schemes ($0<\beta<1$).

For values of the parameter β intermediate between 0 and 1, it was not possible to determine algebraically whether or not the matrix G_∞ could be diagonalized. Thus, to investigate the question a numerical study was done, in which the matrix was constructed for several values of the parameter, and temptatively diagonalized by a call to a routine of the NAG library. When the eigenvalues were

found "evidently distinct", the conclusion was drawn that the matrix was diagonalizable. In numerically non-trivial cases, the matrix X formed of the (possibly generalized) eigenvectors as proposed by the routine, was examined. Here the case of a defective matrix is the limit as $\beta \to \beta^*$ (critical value), of a regular case in which the eigenvectors form a basis. Thus as β approaches the critical value, at least 2 eigenvectors (in practice more than 2) approach the same direction, and coalesce in the limit, resulting in a non-invertible matrix X. This phenomenon was detected by monitoring the condition number κ of the matrix X which suddenly becomes very large in the vicinity of a critical value and is infinite in the limit. Thus, when the matrix C_∞ can be diagonalized, that is,

$$C_\infty = X \Lambda X^{-1} \tag{3.28}$$

in which the matrix Λ is diagonal, we define:

$$\kappa(X) \equiv \| X \|_2 \| X^{-1} \|_2 \tag{3.29}$$

Note that given the matrix C_∞, there is some amount of arbitrariness left in the definition of the matrix X, since the (possibly complex) eigenvectors can be normalized and ordered in many ways. However, if the convention of normalizing the eigenvectors to 1 is made, then $\kappa(X)$ is independent of the choices left free. The reader unfamiliar with this result is referred to Appendix C. This convention was made in all the calculations reported.

Table 1 indicates for different values of the parameter β the corresponding values of the spectral radius, $\rho = \rho(C)$, and of the condition number, $\kappa = \kappa(X)$. The calculations were performed for $N = 10$. It appears that the spectral radius is uniformly approximately equal to 0.5, and nearly symmetrical with respect to $\beta = \frac{1}{2}$. (No particular investigation was carried to determine whether the slight dissymmetry was actual or the fact of numerical inaccuracies, due precisely to the large condition number.) Secondly, the condition number is moderate only for values of β away from both 0 and 1, and can be very large near these limits. The condition number is found symmetrical with respect to $\beta = \frac{1}{2}$ to the accuracy of these estimations.

In conclusion, the analysis suggests that a degradation of the iterative properties of the scheme under study should be observed when operating close to either the pure central scheme limit or the fully upwind scheme limit.

Table 1: Spectral radius and condition number versus β		
β	ρ	κ
0	0.50438	∞
0.05	0.49545	225800
0.25	0.48176	181.1
0.50	0.47553	6.314
0.75	0.48176	181.1
0.95	0.49544	225800
1	0.5	∞

$^{*}N = 10$

3.2. Two-dimensional model problem

One possible two-dimensional extension of Problem (3.1) is given by:

$$\begin{cases} u_t + au_x + bu_y = 0 \quad (a>0,\ b>0) \\ u(x,y,0) = u^0(x,y) \quad (0\leq x\leq L_x, 0\leq y\leq L_y) \\ u(0,y,t) = u(x,0,t) = u_0 \quad (t>0) \end{cases} \tag{3.30}$$

Assume that a uniform rectangular mesh of $N_x \times N_y$ gridpoints is employed so that $\Delta x = L_x/N_x$ and $\Delta y = L_y/N_y$. Two Courant-numbers-like parameters can be defined:

$$\nu_x = \frac{a}{\Delta x}, \quad \nu_y = \frac{b}{\Delta y} \tag{3.31}$$

Also assume that the components of the solution vector U are ordered as follows:

$$U \equiv (U_{1,1}, U_{1,2}, \cdots, U_{1,N_y}, U_{2,1}, U_{2,2}, \cdots, U_{2,N_y}, \cdots, U_{N_x,1}, U_{N_x,2}, \cdots, U_{N_x,N_y})^T \tag{3.32}$$

Then if the term au_x (respectively bu_y) is discretized both explicitly and implicitly as in the previous section, the operator P defining the explicit phase

can now be associated with the following matrix:

$$P = \nu_x \, \delta_{2,x} \otimes I_y + \nu_y \, I_x \otimes \delta_{2,y} \tag{3.33}$$

where

$$\begin{aligned} \delta_{2,x} &= (1-\beta_x)\, \delta_x^c + \beta_x \, \delta_x^u \\ \delta_{2,y} &= (1-\beta_y)\, \delta_y^c + \beta_y \, \delta_y^u \end{aligned} \tag{3.34}$$

In these definitions, the matrices subscripted by x (respectively y) are of dimension $N_x \times N_x$ (respectively $N_y \times N_y$), and the symbol $\otimes$ indicates a Kronecker product so that P is of dimension $N_x N_y \times N_x N_y$. Each matrix being given an appropriate dimension, I_x and I_y are identity matrices, δ_x^c and δ_y^c are central-difference matrices having the structure defined in (3.4), δ_x^u and δ_y^u are second-order backward-difference matrices having the structure defined in (3.5). Similarly, the operator $\widehat{P}$ defining the implicit phase is associated with a matrix having the form:

$$\widehat{P} = \nu_x \, \delta_{1,x} \otimes I_y + \nu_y \, I_x \otimes \delta_{1,y} \tag{3.35}$$

where $\delta_{1,x}$ and $\delta_{1,y}$ have exactly the same structure as δ_1 in (3.7) but are of different dimensions.

For a finite time-step Δt, the amplification matrix is given by (3.8). Therefore as $\Delta t \to \infty$, this matrix tends to

$$C_\infty = I - \widehat{P}^{-1} P \tag{3.36}$$

Clearly, this matrix depends on 3 parameters: β_x and β_y that control the upwinding in the x and y directions respectively, and the ratio ν_y/ν_x. This matrix was evaluated for $N_x = N_y = 5$, and for several values of the 3 parameters. In each case, the spectral radius ρ and the condition number of the eigenvector-matrix κ were calculated numerically.

In a first experiment, the ratio ν_y/ν_x is held fixed to the value 1. Table 2a and Table 2b indicate the behavior of ρ and κ for various values of β_x and β_y. Since $N_x = N_y$, the matrix C_∞ depends symmetrically on β_x and β_y, and therefore only the lower half of these tables are calculated.

Table 2a: Spectral radius as a function of β_x and β_y					
ρ	$\beta_y = 0$	$\beta_y = 0.25$	$\beta_y = 0.50$	$\beta_y = 0.75$	$\beta_y = 1$
$\beta_x = 0$	0.93068	-	-	-	-
$\beta_x = 0.25$	0.80729	0.63215	-	-	-
$\beta_x = 0.50$	0.68881	0.47491	0.40451	-	-
$\beta_x = 0.75$	0.56183	0.34464	0.37199	0.43037	-
$\beta_x = 1$	0.50000	0.38688	0.38857	0.38252	0.50000

$^{*}N_x = N_y = 5$; $\nu_y = \nu_x$

Table 2b: Condition number as a function of β_x and β_y					
κ	$\beta_y = 0$	$\beta_y = 0.25$	$\beta_y = 0.50$	$\beta_y = 0.75$	$\beta_y = 1$
$\beta_x = 0$	∞	-	-	-	-
$\beta_x = 0.25$	154.2	536.3	-	-	-
$\beta_x = 0.50$	602.3	218.3	138.3	-	-
$\beta_x = 0.75$	3946	345.8	246.7	1388	-
$\beta_x = 1$	∞	∞	∞	∞	∞

$^{*}N_x = N_y = 5$; $\nu_y = \nu_x$

The spectral radius is always found close to ½ except when using central-differencing in at least one direction (β_y or $\beta_x = 0$). The minimum is not achieved for $\beta_x = \beta_y = ½$, but the value obtained at this point is one of the smallest ones. Examining Table 2b, note that for $\beta_x = \beta_y = 0$ (central scheme in both directions), the software routine produces a proposed eigenvector matrix X, but fails to evaluate the condition number, the largest eigenvalue of $X^{*}X$ being close to 12 while the smallest one is found very small but negative (close to -4.10^{-14}) which is mathematically incorrect. For cases where the fully upwind scheme is

used in at least one direction (β_y or $\beta_x = 1$) the software routine fails to produce an eigenvector matrix. For all these pathological cases noticeable in Table 2b by $\kappa = \infty$, the amplification matrix G_∞ is strongly suspected to be defective.

In a second experiment, the upwinding parameters β_x and β_y are held fixed to the value ½, and the ratio ν_y/ν_x is the variable parameter. Since the matrix G_∞ is unchanged when the value of the parameter is changed in its inverse, that is when the roles of x and y are permuted, this ratio is only assigned values between 0 and 1.The results of this experiment are indicated in Table 3.

Table 3: Spectral radius and condition number versus ν_y/ν_x

ν_y/ν_x	ρ	κ
0	0.40451	10.59
0.01	0.40451	153.9
0.20	0.40451	131.3
0.40	0.40451	126.3
0.60	0.40451	124.7
0.80	0.40451	124.1
0.99	0.40451	124.0
1	0.40451	138.3

$^{\bullet}N_x = N_y = 5\ ;\ \beta_x = \beta_y = ½$

The same value of the spectral radius is found in all cases while the condition number depends only very weakly on the parameter except that it is discontinuous at both limits of the range. (This latter point is not surprising, because at each limit the amplification matrix G_∞ must have multiple eigenvalues: (1) for $\nu_y = 0$, because we have a repetition of N_y identical one-dimensional problems, and (2) for $\nu_y = \nu_x$ due to the symmetry, we observe several double eigenvalues.) We conclude from this experiment, that using an average amount of upwinding in both directions, that is $\beta_x = \beta_y = ½$, results in a satisfactory scheme from the iterative convergence standpoint even when the wavespeeds in the two directions are very different.

In conclusion of this section, it appears that in two dimensions also, the numerical scheme is more efficient when the upwinding parameters β_x and β_y are set equal to the average value ½, than when operating close to either the central-differencing or the fully-upwind differencing limits.

3.3. Solution of the one-dimensional model problem.

In this section, we present a series of numerical experiments related to the wave equation. The solution to the problem defined in (3.1) was calculated numerically by the various schemes defined in the previous section. The constant u_0 was set equal to 0 and so the steady-state solution was identically equal to 0 and the components of the vector U^n were pure error to be dissipated. Three types of initial solution have been used:

(a) $U^0 = W$, where W is the vector of (3.22)

(b) $U^0 = (1,0,0,\ldots,0)^T$, referred to as a Dirac on the right side, and

(c) $U_j^0 = f(j)$ where the function f was a built-in function whose values are random numbers in the range $[-1,1]$.

The initial solution given in (a) is interesting since it is the highest-frequency mode thus the mode the least dissipated by the schemes studied in the previous subsection, while in the solution given in (c) all the modes are excited. The solution given in (b) was examined because the convergence can be studied analytically in the case of the fully-upwind second-order scheme.

For each experiment, a figure shows the variation with iterations n, of the norm $\| U^n \|_\infty$, plotted on a $\log_{10}$-scale and indicated by a solid line. Since many of the schemes presented in (3.1) are associated with an amplification matrix G, having a spectral radius equal to or close to ½, the sequence 2^{-n} is also indicated on these plots by a dashed line.

Figures 2-8 are related to the central-differencing scheme defined in (3.18). On Figure 2, the initial solution is the highest-frequency mode, the most difficult case. With iterations, the norm $\| U^n \|_\infty$ increases to the value 2, but never goes beyond this limit since, by virtue of (3.23),

$$\begin{aligned} \forall n \geq 2, \quad \| U^n \|_\infty &= \| (G_\infty^c)^n \, U^0 \|_\infty \\ &\leq \| (G_\infty^c)^n \|_\infty \, \| U^0 \|_\infty = (2-2^{-n}).1 \\ &< 2 \end{aligned} \tag{3.37}$$

Then it takes about 200 iterations, that is $2N$, before the application of the

iterative scheme has the apparent effect of reducing the norm. It also appears that the line that represents the sequence 2^{-n} on this semi-log plot is an asymptotic direction for the sequence of norms, but not a true asymptote. This confirms exactly results proved in the previous subsection.

In the case of Figure 3, the initial solution is a Dirac on the right side. The figure shows that certain modes, this one in particular, are indeed attenuated at the rate of 2^{-n}.

In the case of Figure 4, the initial solution is made of random numbers. In this way, all the modes are initially excited in the error vector. The history plot is of course found less regular, but the general trend is analogous, exhibiting the same pathology as in the worst case (Figure 2).

The cases of Figures 5-6 are analogous to the cases of Figures 2 and 4, except they correspond to a smaller mesh, $N = 50$. There it appears that the break between the "pseudo-stationary" segment and the "dissipative-convergence" phase occurs at about $n = 100$, that is again $2N$. This confirms the formula that predicts that the break should occur at about $n = \frac{N}{1-\rho}$ since here, $\rho = \frac{1}{2}$.

Figures 7-8 indicate how the convergence history becomes more regular when artificial dissipation is added, which is not really new. To realize these experiments, the partial-differential equation was modified to

$$u_t + c\, u_x = \varepsilon\, \Delta x^2\, u_{xx} \tag{3.38}$$

and while the term u_x was represented as before by $\delta_2 U^n / \Delta x$, the additional term $\varepsilon\, \Delta x^2\, u_{xx}$ was approximated by $\varepsilon Trid(1,-2,1)U^n$. In this way the artificial dissipation is of the same order as the formal truncation error at the steady state, provided ε is of order 1, and preferrably a fraction of 1 for accuracy. Therefore the amount of artificial dissipation in the case of Figure 7 is already substantial while it is excessive in the case of Figure 8 (-note the little importance of the spectral radius of the amplification matrix in the convergence process-). This illustrates, as well known, that artificial dissipation can improve the convergence of central schemes but the required amount can destroy the nature of the hyperbolic problem, unless a sophisticated evaluation of the dissipation term is devised, which was not done here.

We now turn to the fully-upwind second-order scheme. Figures 9-13 are analogous to Figures 2-6, the only difference is the scheme employed.

Figure 9 demonstrates that the norm $\| U^n \|_\infty$ is unchanged through the first $2N$ iterations, and only then it rapidly converges. $\mathrm{Log}(2^{-n})$ is again an

asymptotic direction. The norm never goes beyond 1, since by virtue of (3.27),

$$\forall\, n \geq 2, \quad \| U^n \|_\infty = \| (G_\infty^u)^n\, U^0 \|_\infty$$

$$\leq \| (G_\infty^u)^n \|_\infty \, \| U^0 \|_\infty = 1 \tag{3.39}$$

When a Dirac on the right side is imposed as initial solution (Figure 10), the convergence of $\| U^n \|_\infty$ can be shown to initially be like $\sqrt{\frac{2}{\pi n}}$ and this is confirmed by the calculation.

For an initial distribution of errors made of random numbers (Figure 11), the convergence is found monotonic; the plot of the norm shows one angular point corresponding to a change in the index j for which $\| U^n \|_\infty = |U_j^n|$.

Figures 12-13 correspond to cases where a smaller mesh is employed ($N = 50$). Note again that the initial phase of "pseudo-stationary" convergence extends to about $2N$ iterations.

We now evaluate numerically the half-upwind scheme ($\beta = \frac{1}{2}$).

Figures 14-15 demonstrate for a particular, and for a general initial solution that the pathology in convergence observed in the previous two schemes is eliminated. The solution converges globally at the same rate as 2^{-n} (or slightly faster).

Essentially the same convergence plots are obtained for the third-order scheme ($\beta = 1/3$). This demonstrates that it is for a whole range of values of β about $\frac{1}{2}$ that the convergence of the scheme is satisfactory.

In conclusion, these experiments confirm that the convergence of both the central-differencing scheme ($\beta = 0$) and the fully-upwind second-order scheme ($\beta = 1$) are pathological. In contrast, the half-upwind scheme ($\beta = \frac{1}{2}$) and schemes corresponding to an upwinding parameter β in a range about $\frac{1}{2}$ have a satisfactory convergence.

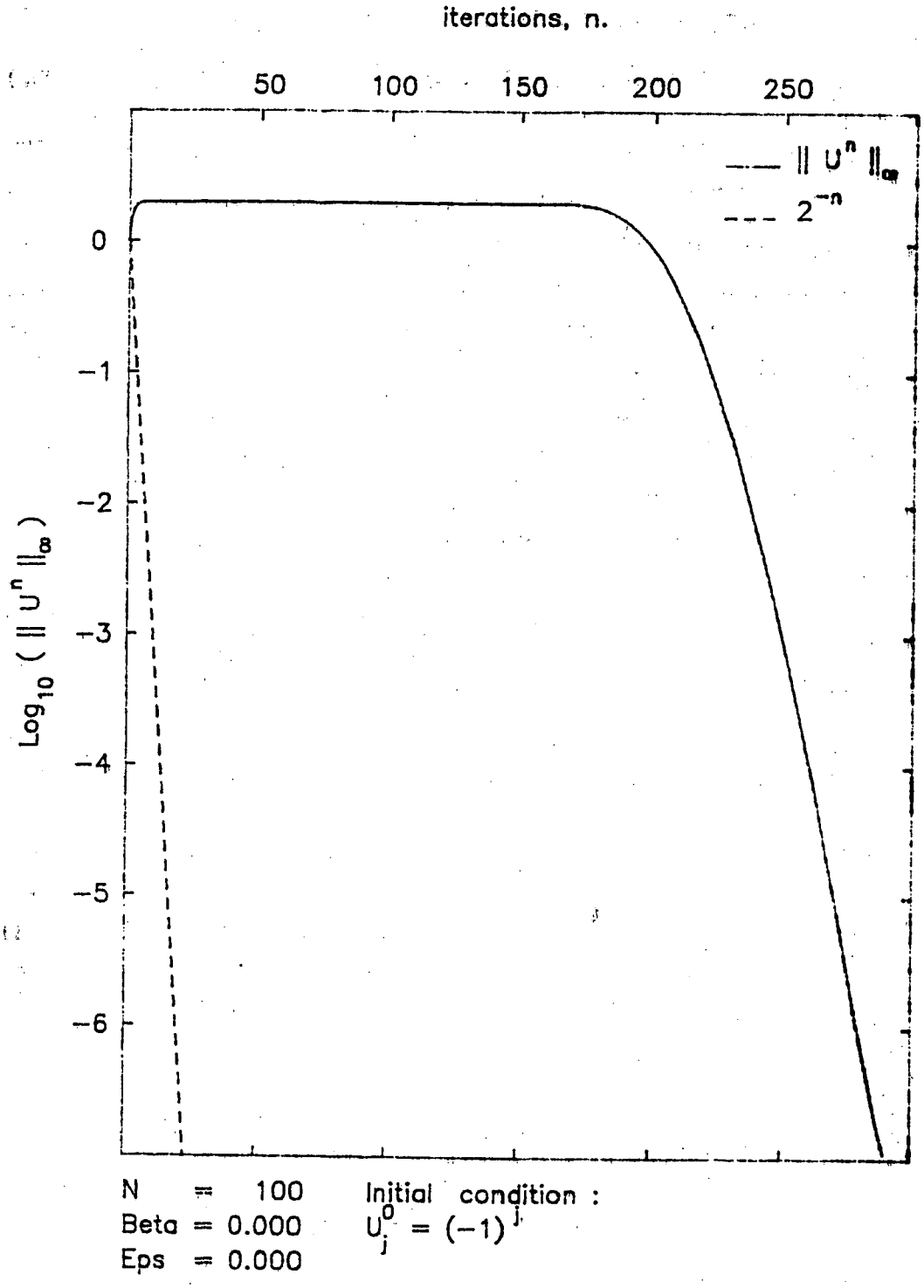

Figure 2. Central–differencing scheme. Case where the initial solution is the highest–frequency mode.

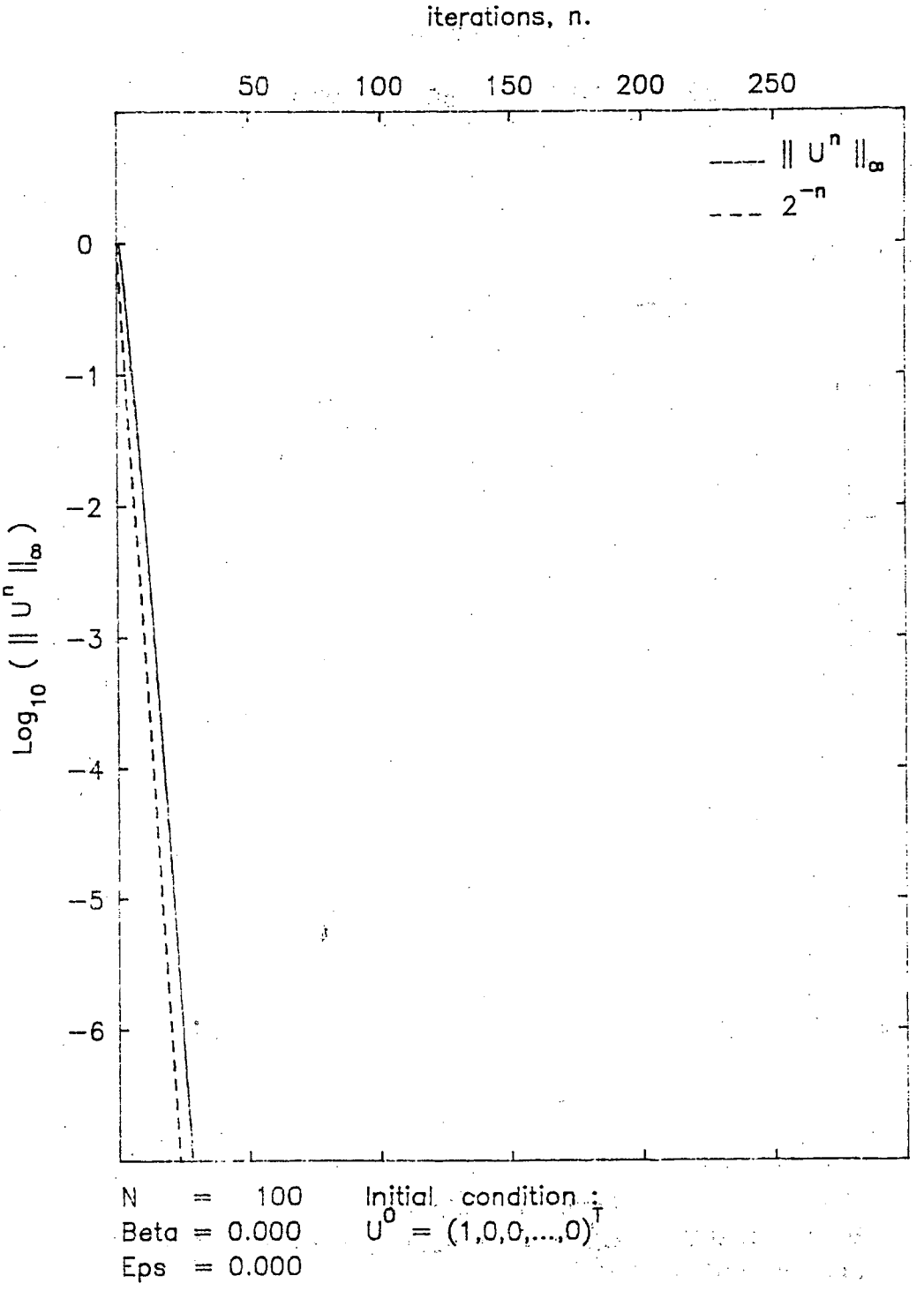

Figure 3. Central–differencing scheme. Case where the initial solution is a Dirac on the right side.

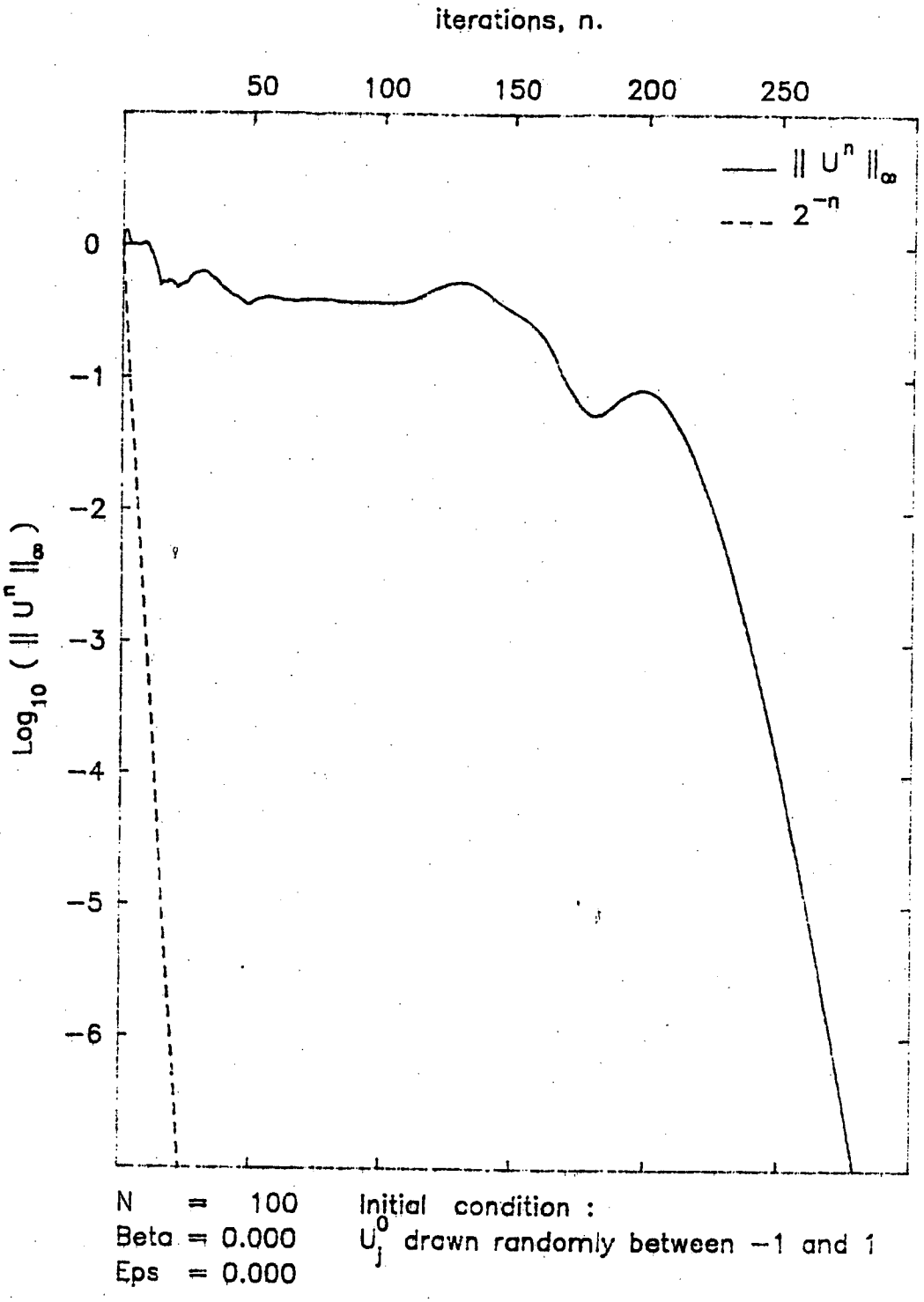

Figure 4. Central–differencing scheme.
Case where the components of the initial solution are random numbers.

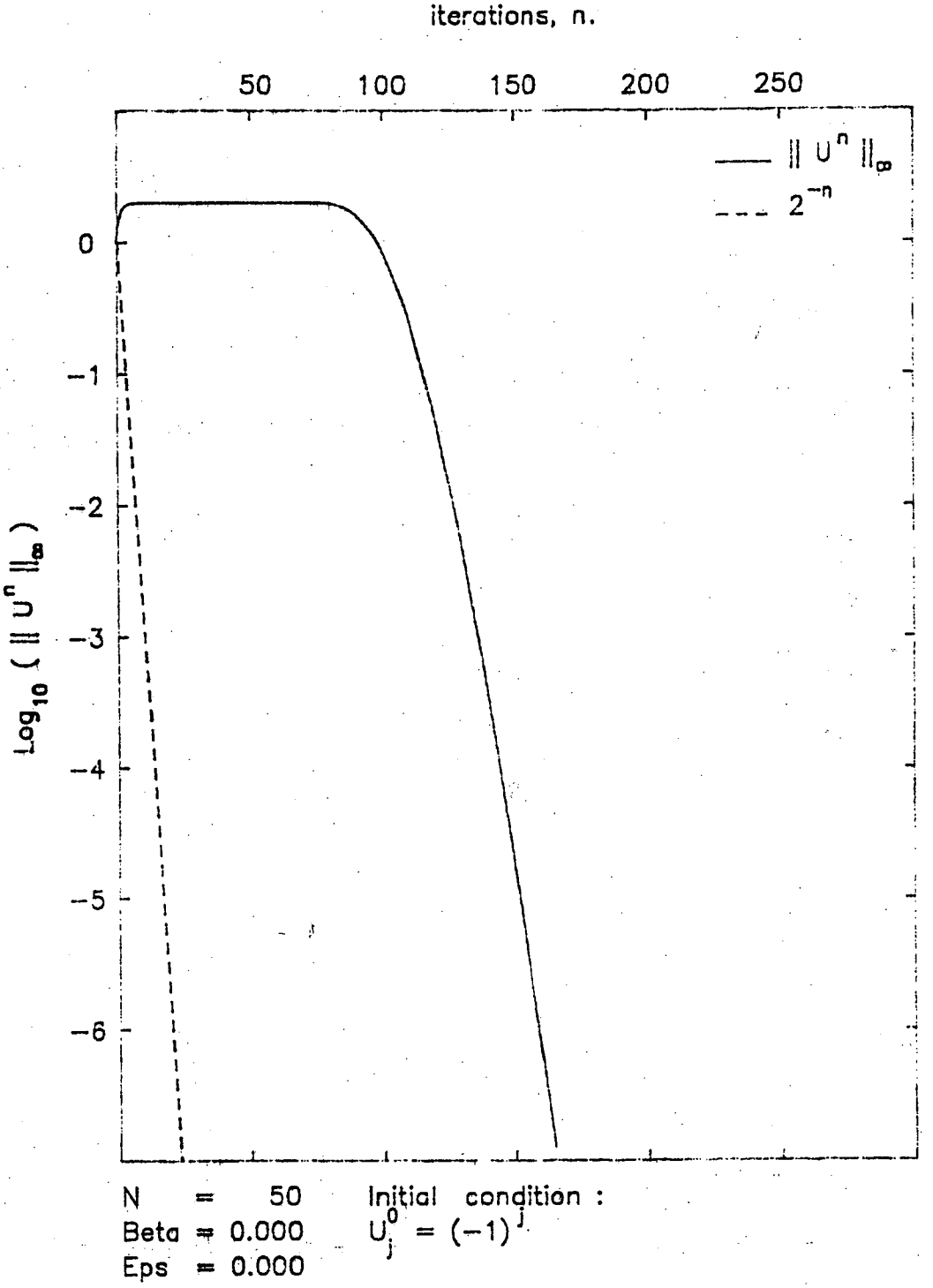

Figure 5. Central–differencing scheme.
A smaller mesh.
Case where the initial solution is the highest–frequency mode.

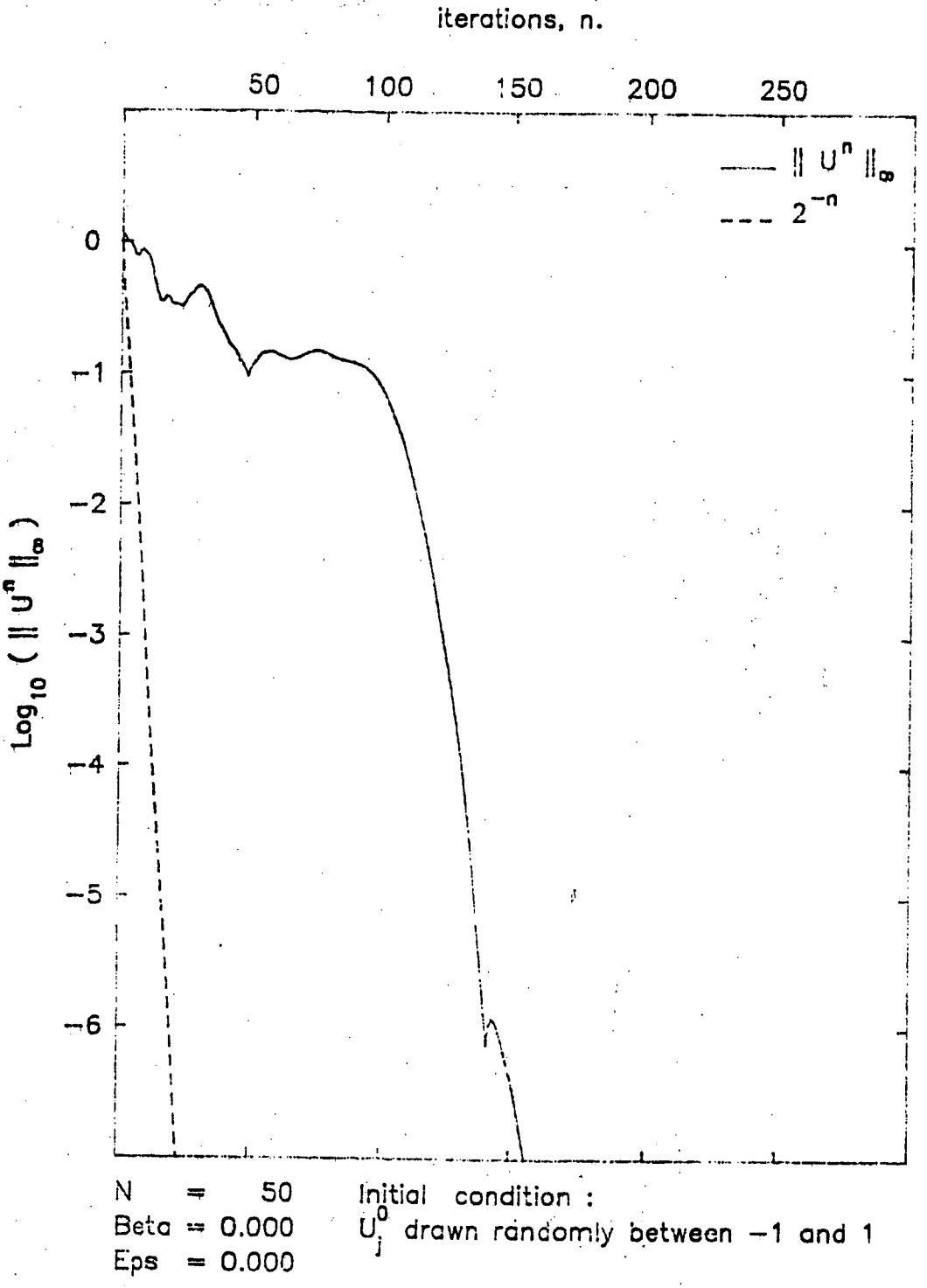

Figure 6. Central–differencing scheme.
A smaller mesh.
Case where the components of the initial solution are random numbers.

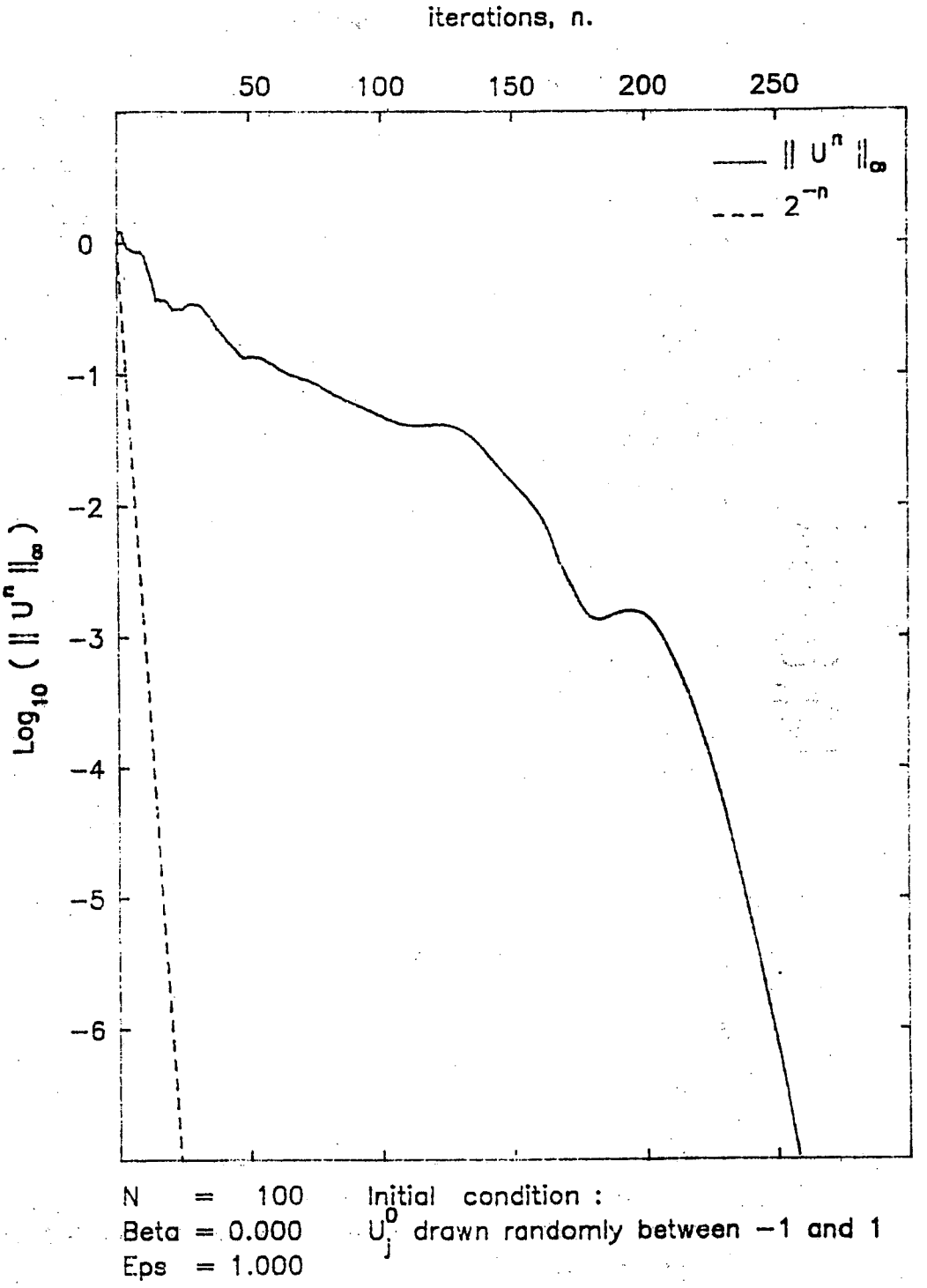

Figure 7. Central–differencing scheme.
The effect of a fair amount of artificial dissipation.
Case where the components of the initial solution are random numbers.

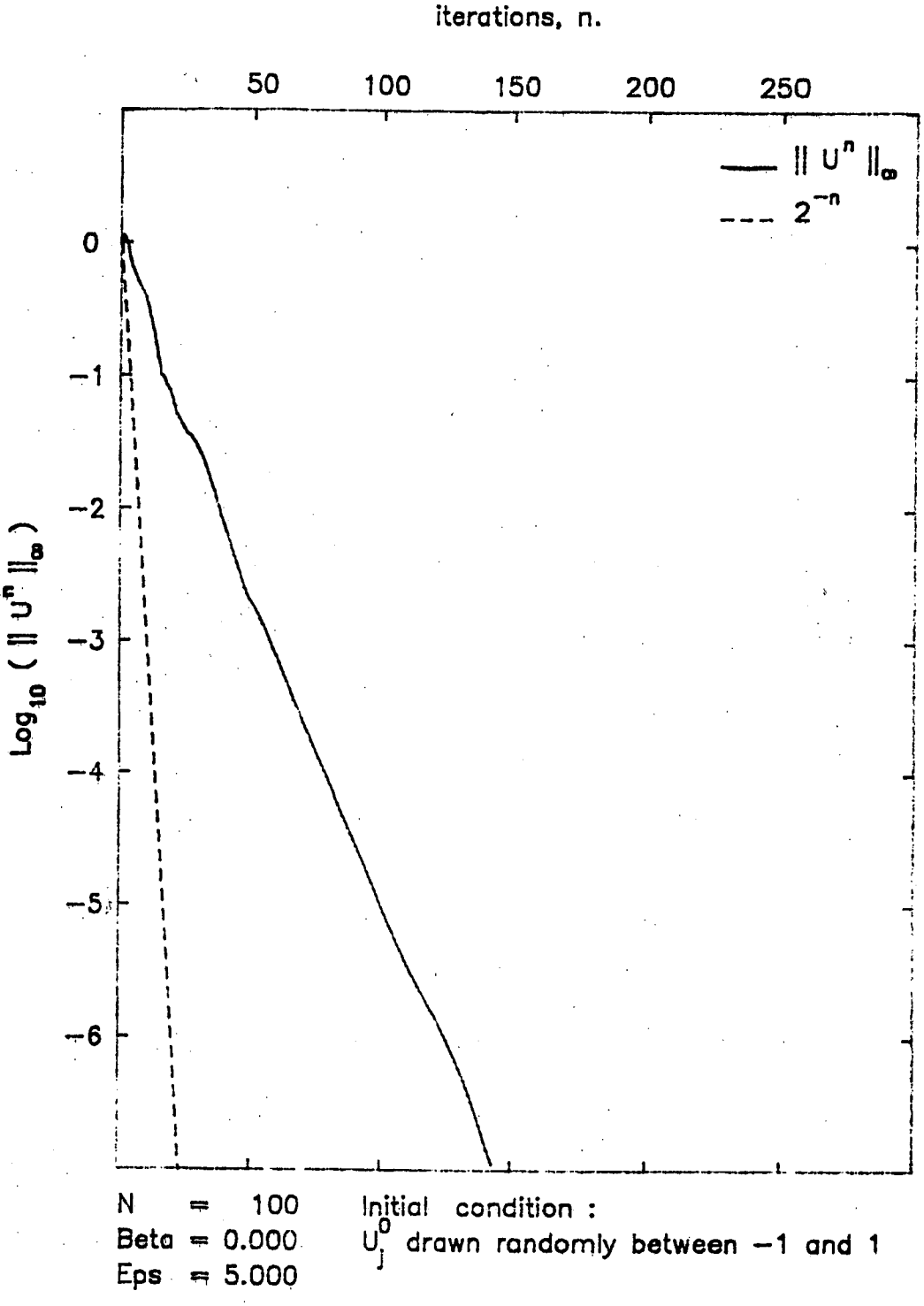

Figure 8. Central–differencing scheme.
The effect of a very large amount of artificial dissipation.
Case where the components of the initial solution are random numbers.

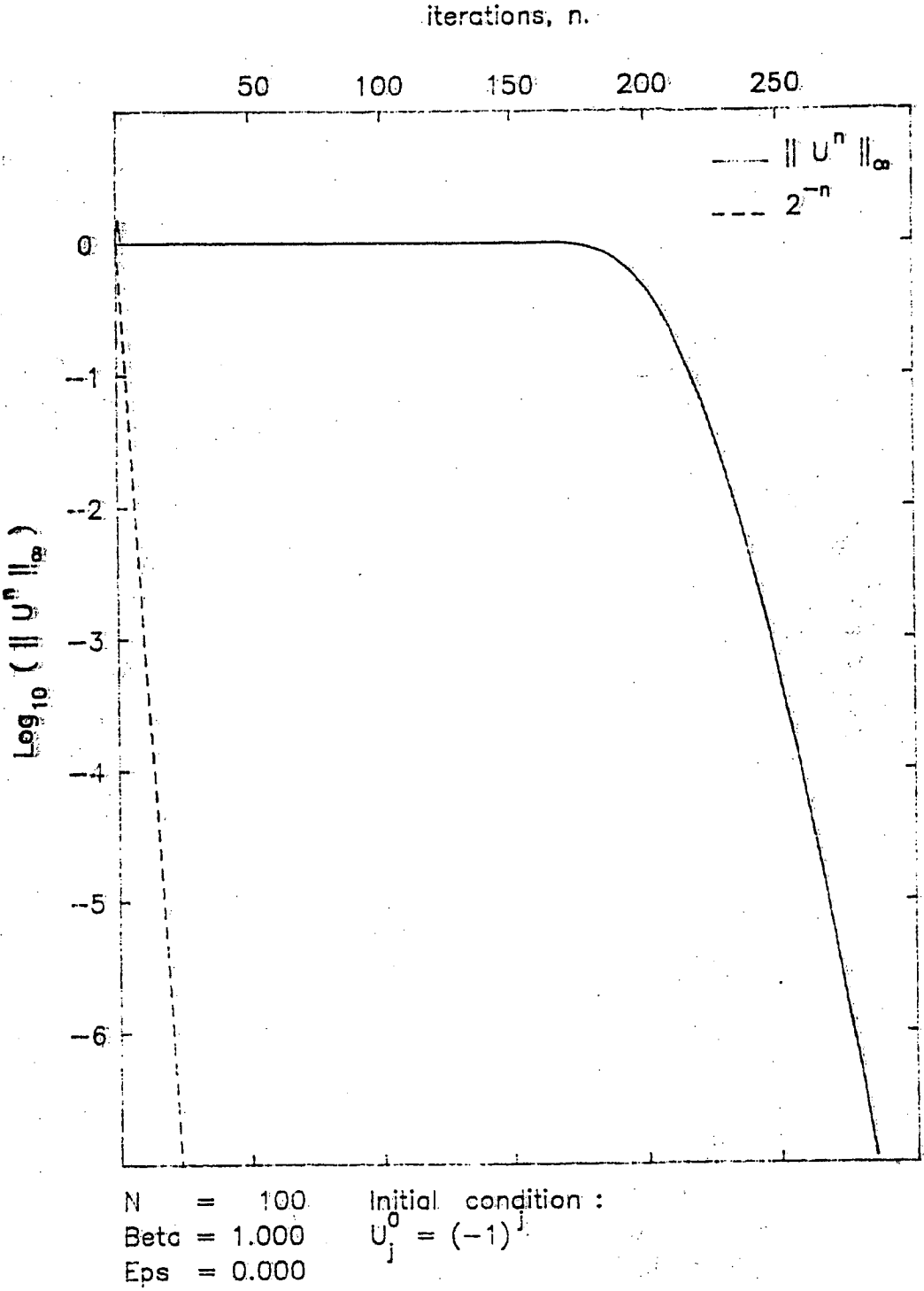

Figure 9. Fully-upwind second-order scheme. Case where the initial solution is the highest-frequency mode.

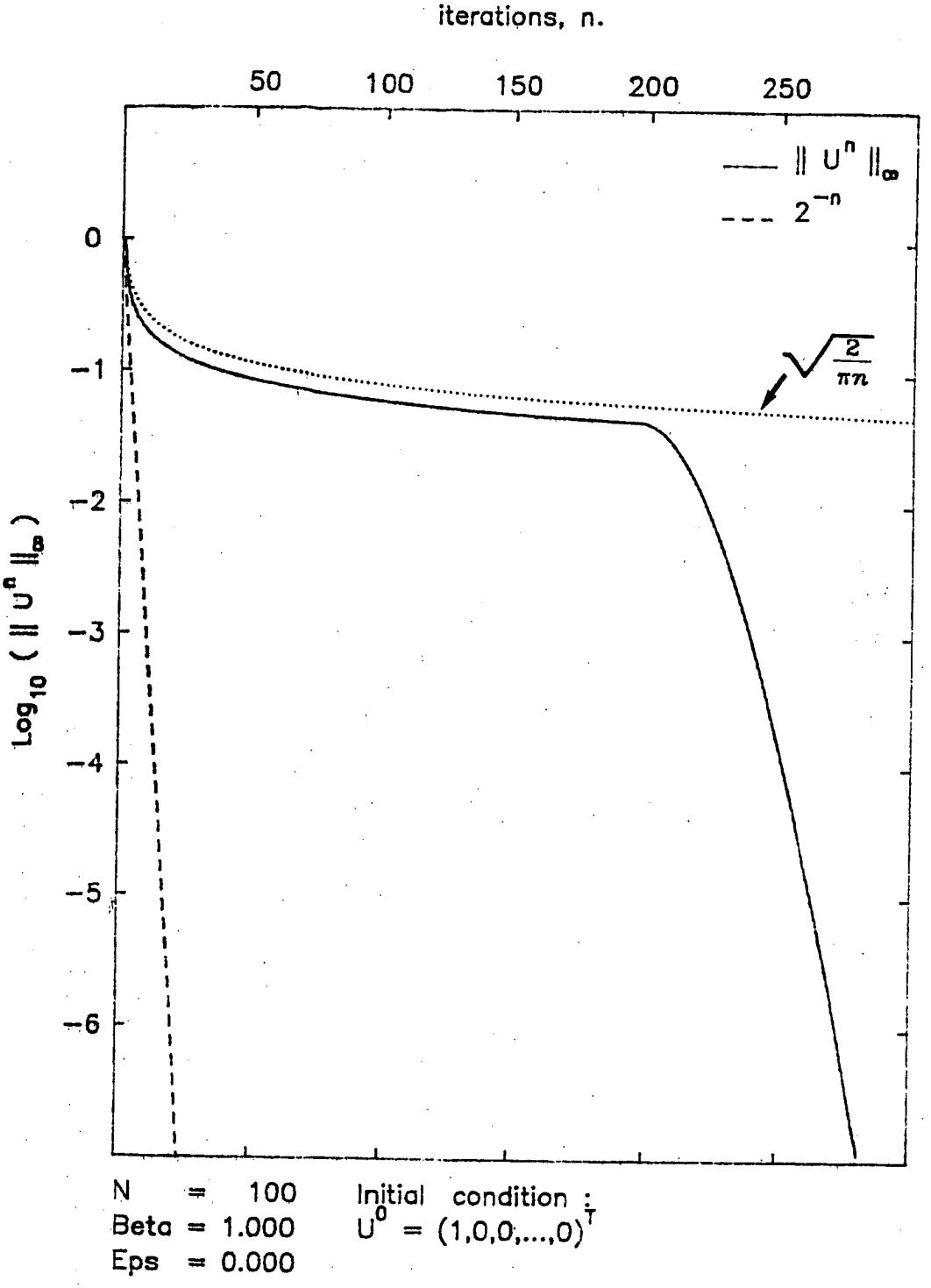

Figure 10. Fully–upwind second–order scheme. Case where the initial solution is a Dirac on the right side.

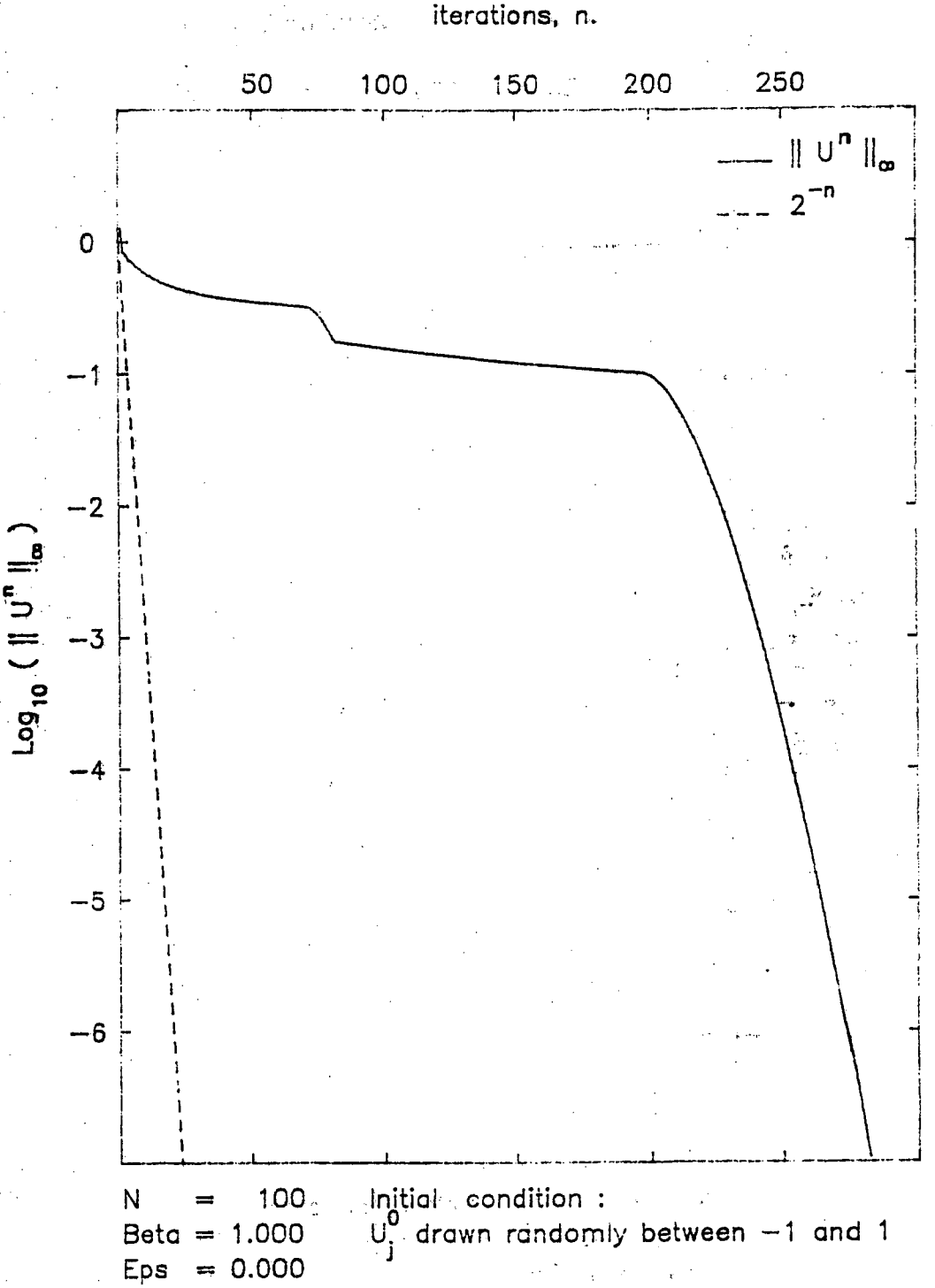

Figure 11. Fully-upwind second-order scheme. Case where the components of the initial solution are random numbers.

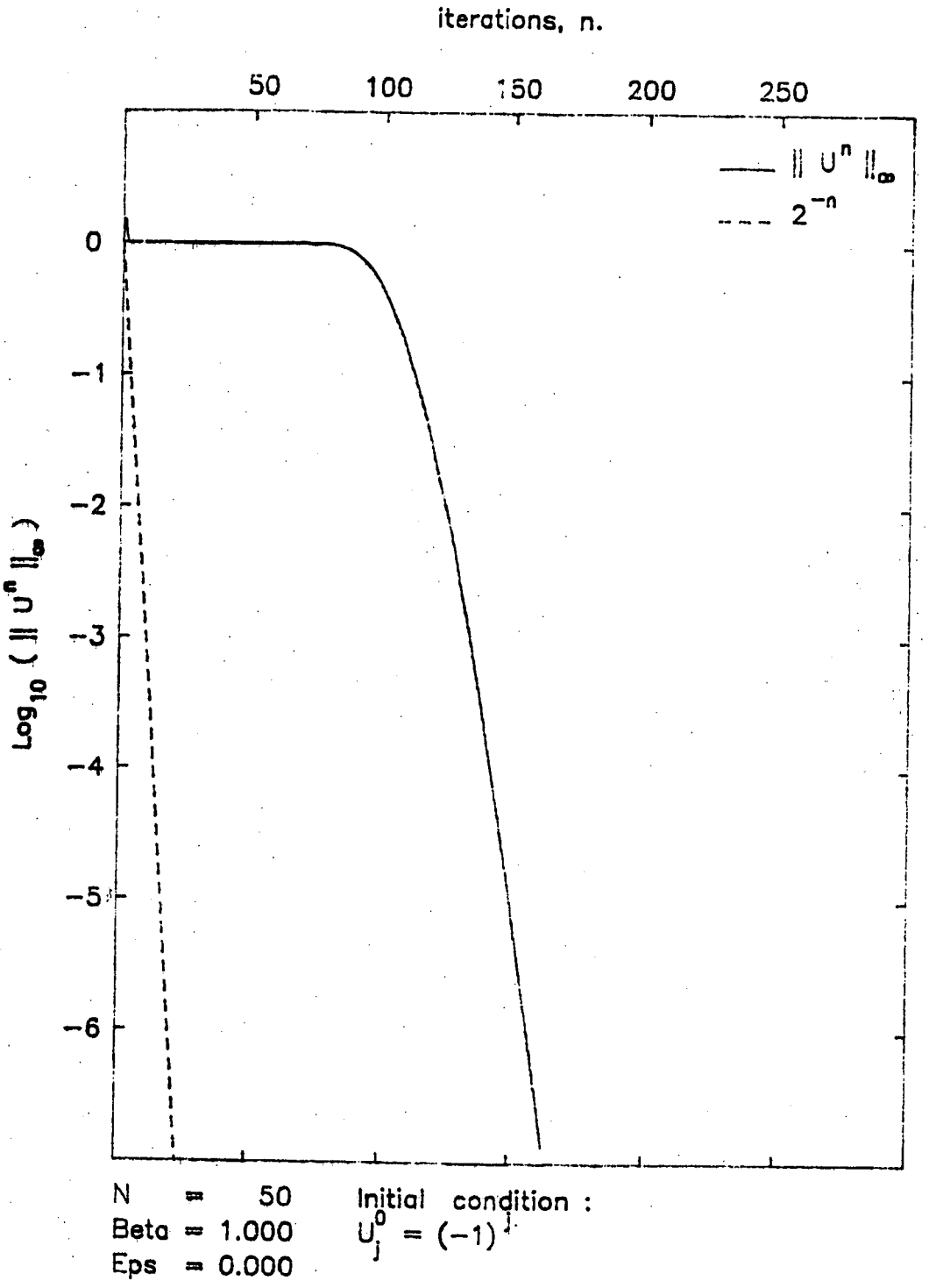

Figure 12. Fully–upwind second–order scheme.
A smaller mesh.
Case where the initial solution is
the highest–frequency mode.

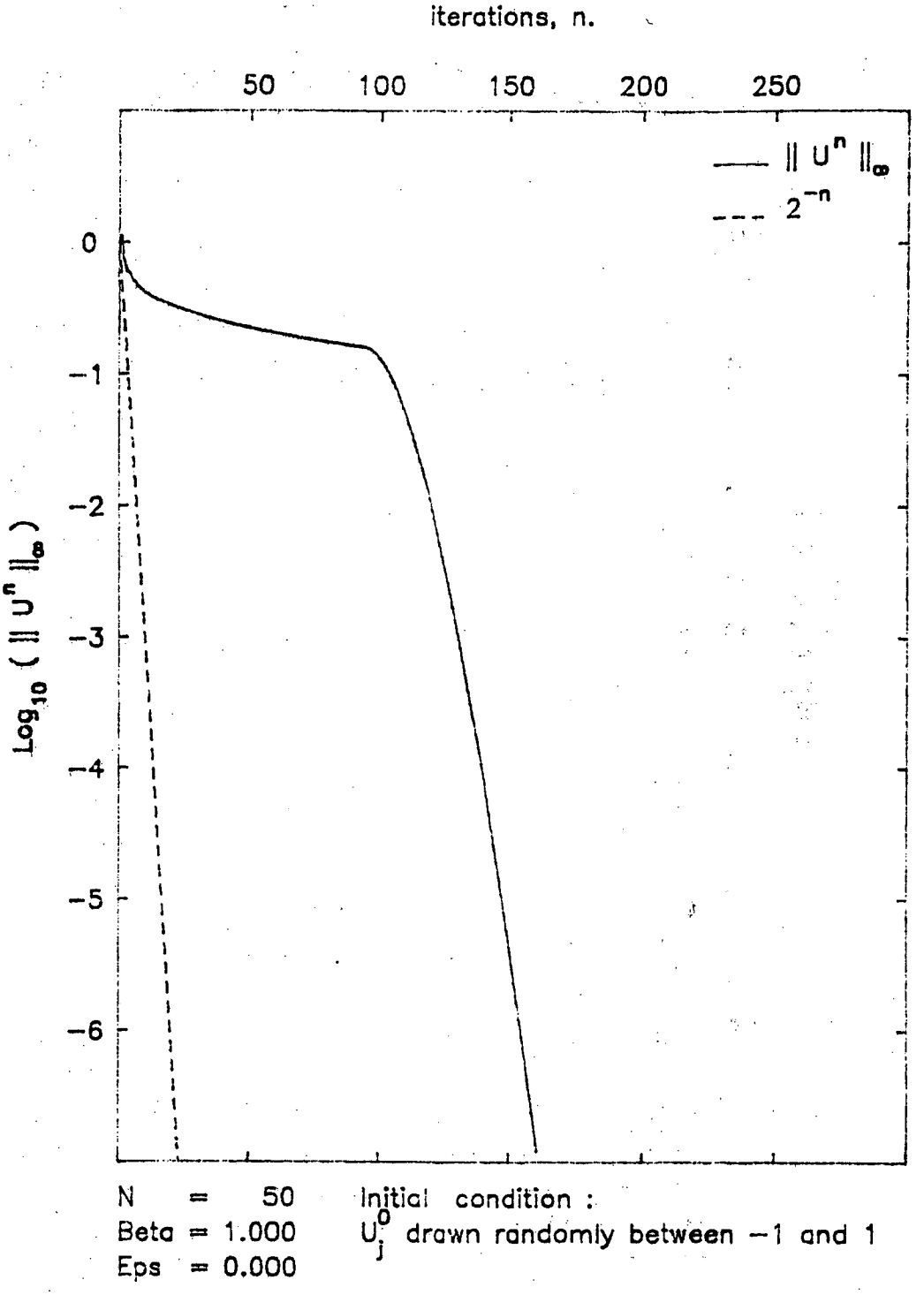

Figure 13. Fully–upwind second–order scheme.
A smaller mesh.
Case where the components of the initial solution are random numbers.

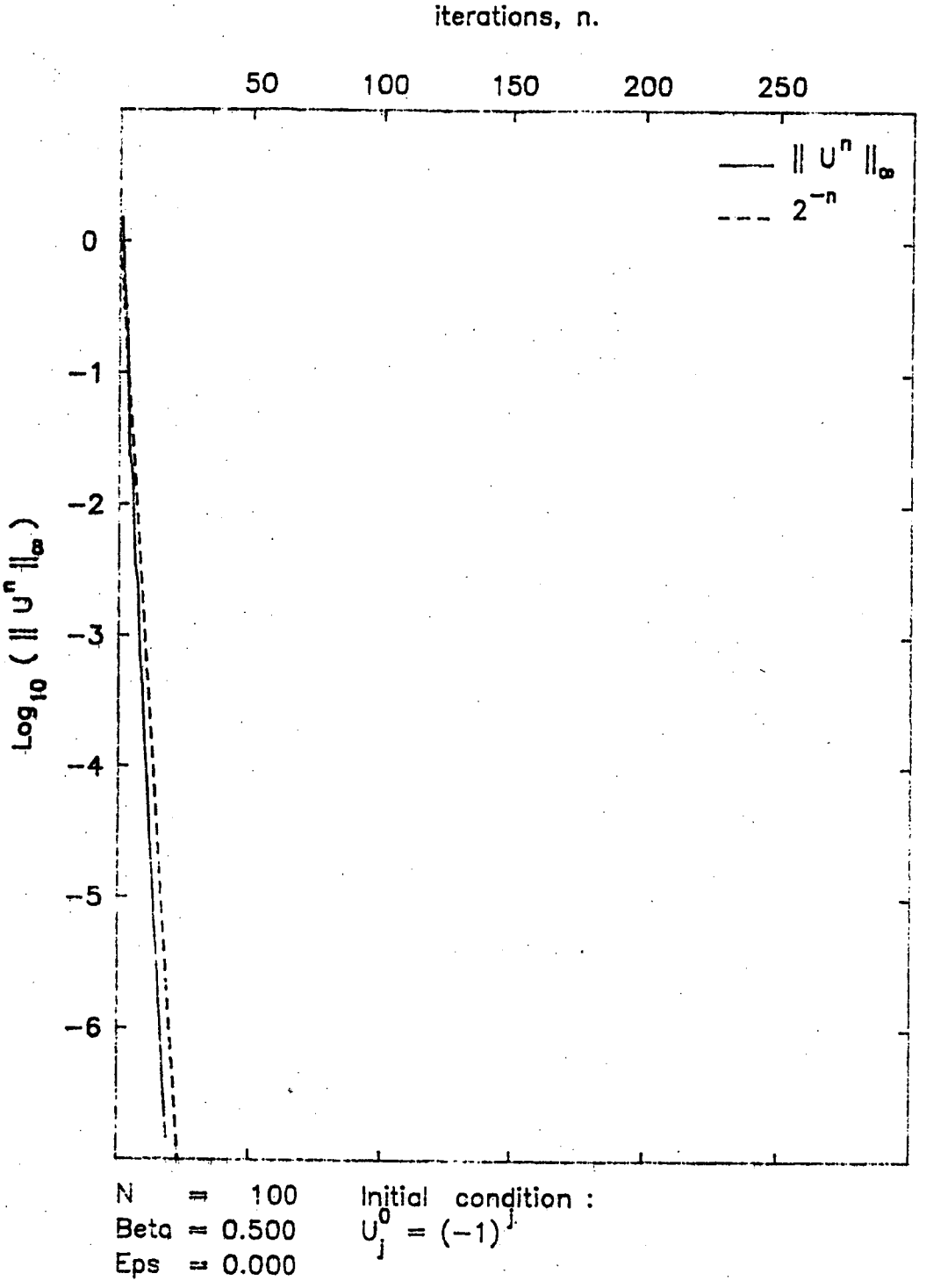

Figure 14. Half-upwind scheme.
Case where the initial solution is
the highest-frequency mode.

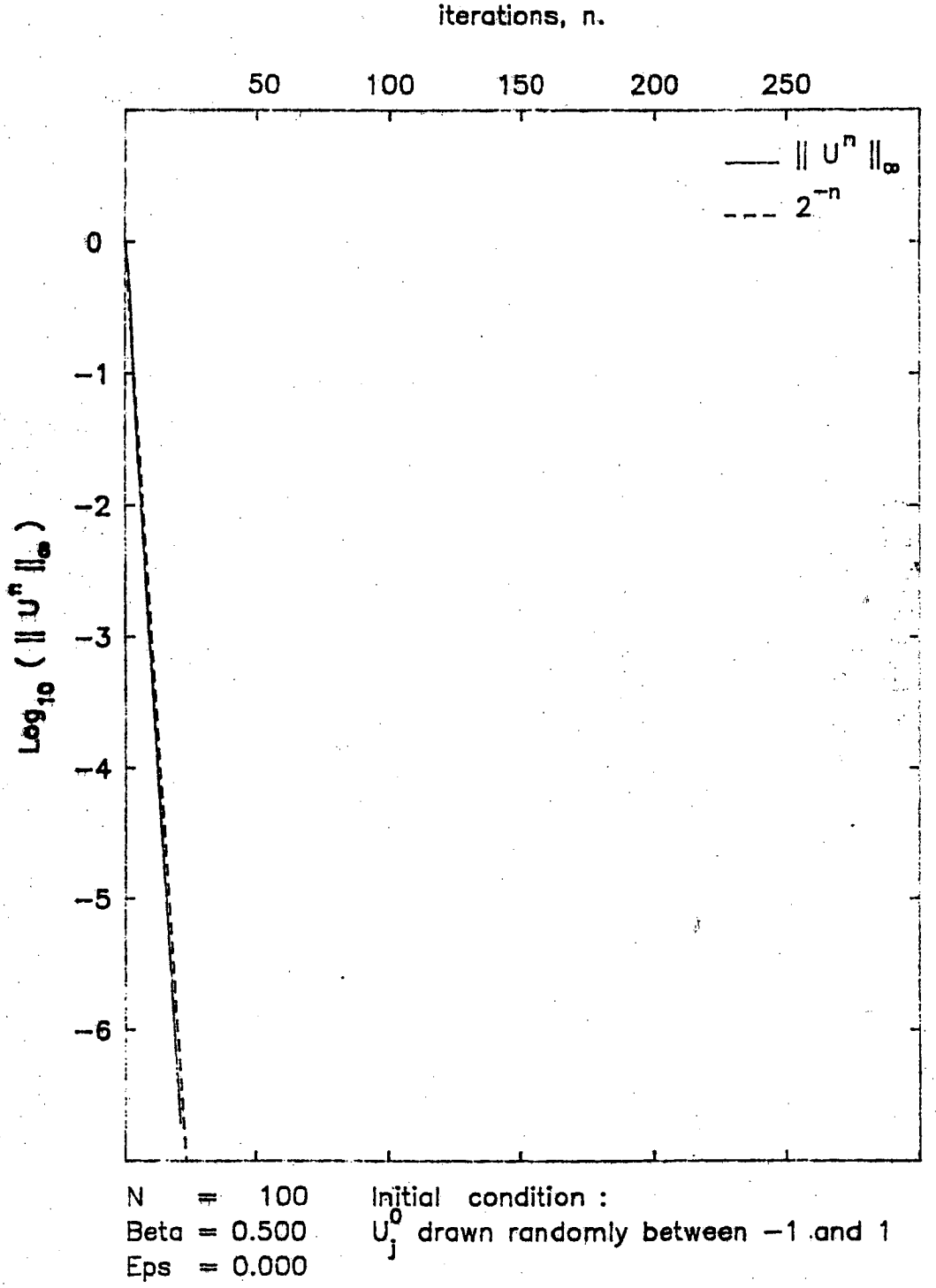

Figure 15. Half-upwind scheme.
Case where the components of the initial solution are random numbers.

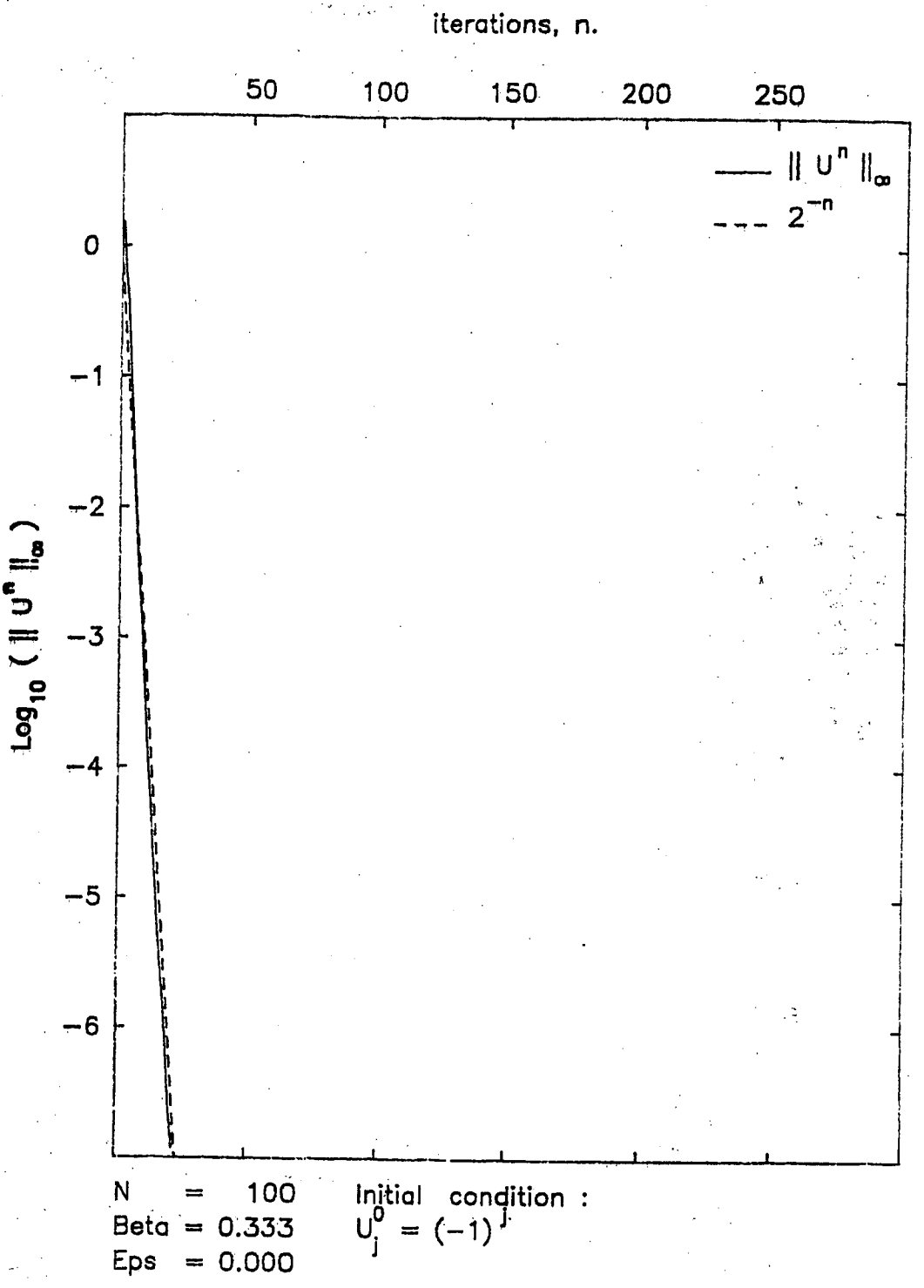

Figure 16. Third–order scheme.
Case where the initial solution is
the highest–frequency mode.

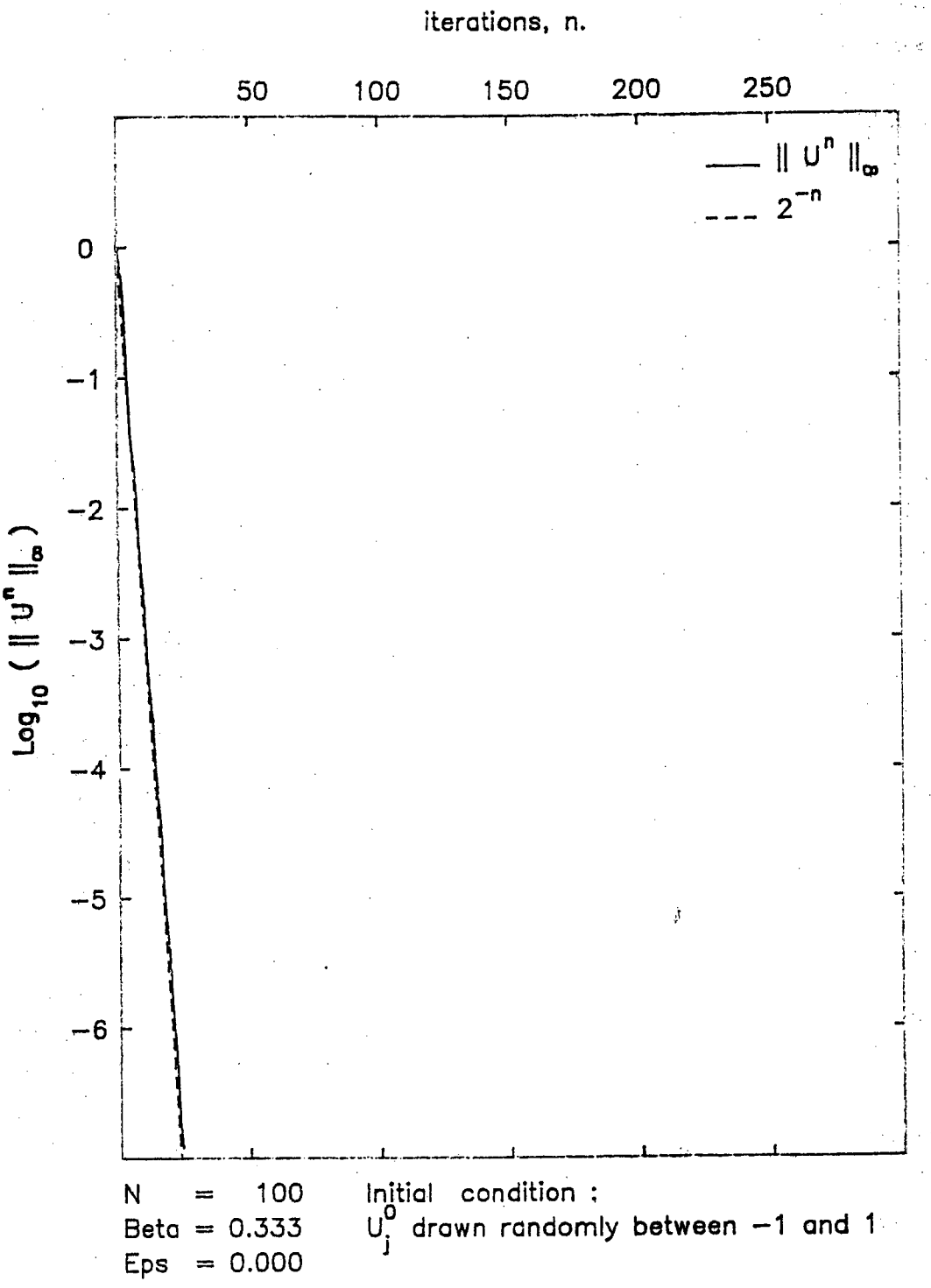

Figure 17. Third-order scheme.
Case where the components of the initial solution are random numbers.

4. NUMERICAL EXPERIMENTS ON THE 2-D EULER EQUATIONS

In order to evaluate some of the effects on iterative convergence of the degree of upwinding in problems where the Euler equations are solved, numerical experiments were made on a simple two-dimensional problem.

We considered the flow in a channel over a 4.2% thick circular bump. The algorithm was based on a finite-element formulation using P1-Lagrange interpolation. A rather coarse mesh of 161 points shown on Figure 18, was employed.

Firstly a central-differencing approximation was constructed using a Galerkin approach. Secondly, an upwind-differencing approximation was constructed based on a flux-splitting procedure that will be described elsewhere [13]. The two differencing approximations were then combined linearly in a way similar to that of (3.3), and a parameter β again controlled the degree of upwinding introduced in the differencing ($\beta = 0$: central scheme; $\beta = 1$: fully upwind scheme). The combination constituted the explicit phase. Implicitly, a first-order approximation was used. However, the implicit factor was not solved completely, but only partially by 2 relaxation sweeps. This is one of several points by which the simulation departed from the theoretical model.

The solution was impulsively started from uniform flow. Therefore, contrasting with the linear model problem, there was there, inevitably, an initial phase in the convergence history during which the time integration captured a truly transient solution. For this reason in particular and because the formulation was nonlinear, it was not possible to utilize infinite time-steps, the transient phase being usually particularly subject to numerical instabilities. Thus, in order to mimic the theoretical situation studied in analysis, the time-step Δt was updated at each new iteration so that

$$C.F.L.(n\Delta t) = n \qquad (4.1)$$

where n is the iteration counter.

These decisions being made, a few preliminary runs were made. They revealed that the fully upwind scheme was unstable. Therefore, in order to obtain a stable scheme partially preserving monotony we introduced limiters in the subsequent experiments.

On Figure 19 is indicated the convergence history for different values of the upwinding parameter β for 3 different flow regimes.

In the case of a purely subsonic flow, $M_\infty = 0.50$ (Figure 19a), it can be seen that the central-differencing scheme ($\beta = 0$) performs evidently more poorly than the other schemes. As β increases, the convergence becomes more rapid.

Suddenly, for $\beta = 1$, the scheme becomes unstable after 75 iterations and no longer converges if Δt is further increased. On this plot, a dashed line indicates the convergence history of the first-order scheme that is obtained when using in the explicit phase the same first-order approximation as in the implicit phase. This scheme which approaches Newton's method as $\Delta t \to \infty$ (or $n \to \infty$), was found faster as expected, but of course not as accurate. We mainly conclude from this experiment that the fully upwind scheme is not adequate in the limit of an infinite time-step.

In the case of a transonic flow, $M_\infty = 0.85$ (Figure 19b), the central-differencing scheme was found unstable and also the scheme obtained for $\beta = 1/10$. As β increases, the convergence to steady state again becomes more rapid except when β achieves the value 1. The fully upwind scheme for large time-steps seems to enter a limit cycle. This may be due to the use of limiters and non-differentiable flux-splitting. We retain that this phenomenon does not occur for other values of β. Hence we conclude again that the fully upwind scheme is pathological. Again the first-order scheme is found to be the fastest.

Finally, a purely supersonic flow was computed corresponding to $M_\infty = 1.50$ (Figure 19c). In this case, the central-differencing scheme was again found to be unstable. As β increases and approaches ½ the convergence becomes more rapid. But if β is increased further the convergence is not as fast. Here, the fully upwind scheme is not found pathological. However, it is rather paradoxical that for this flow that admits a preferential direction and could be solved by a space-marching integration procedure, that the fully upwind scheme is not the most efficient scheme, the maximum convergence rate being achieved by the half-upwind scheme. Surprisingly, the first-order scheme was found less efficient. No precise reason is known to explain this.

We conclude this section by noting that we could not realize in the nonlinear case the precise conditions of the theoretical analysis: nonlinearity, use of limiters and non-differentiable flux-vector splitting, incomplete inversion of the implicit factor, finite time-step. Thus it is not surprising that the numerical schemes could not show exactly the same behaviour. However, in the nonlinear case as in the linear case, the implicit scheme based on the fully-upwind differencing revealed to have certain undesirable convergence anomalies.

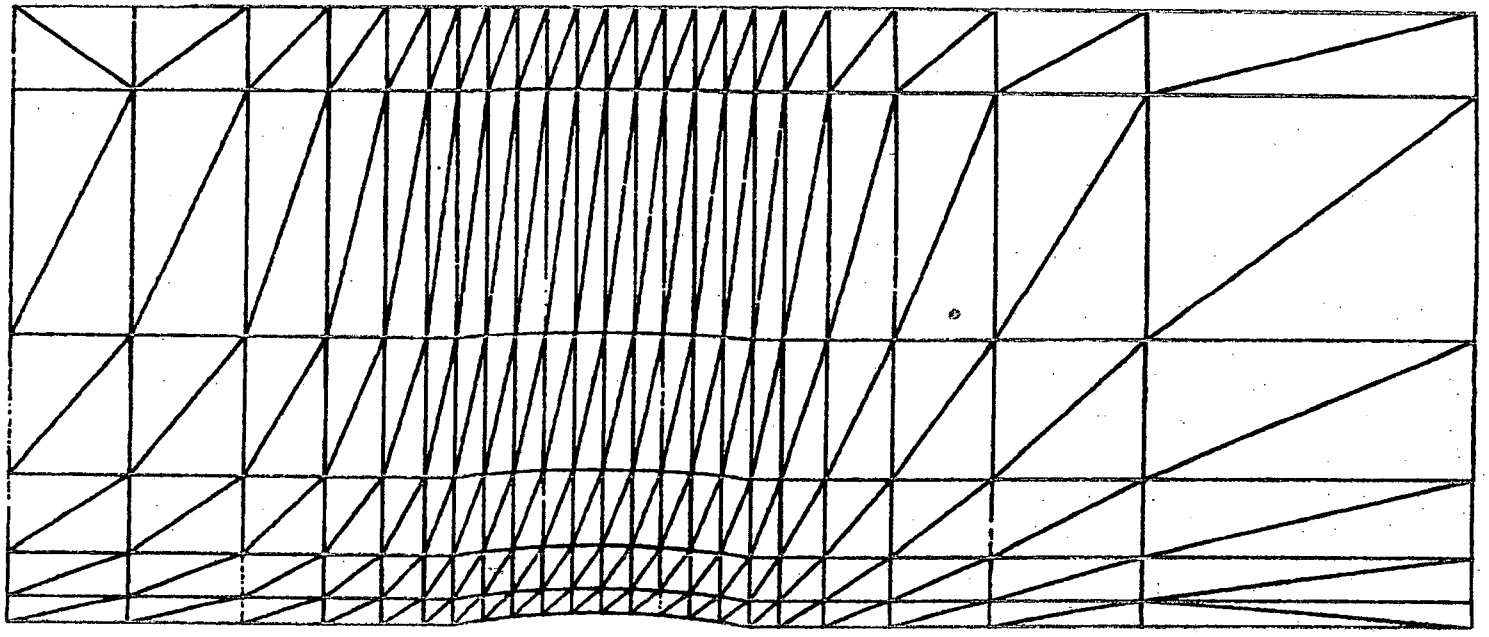

Figure 18. Finite-element mesh employed in the Euler simulations.

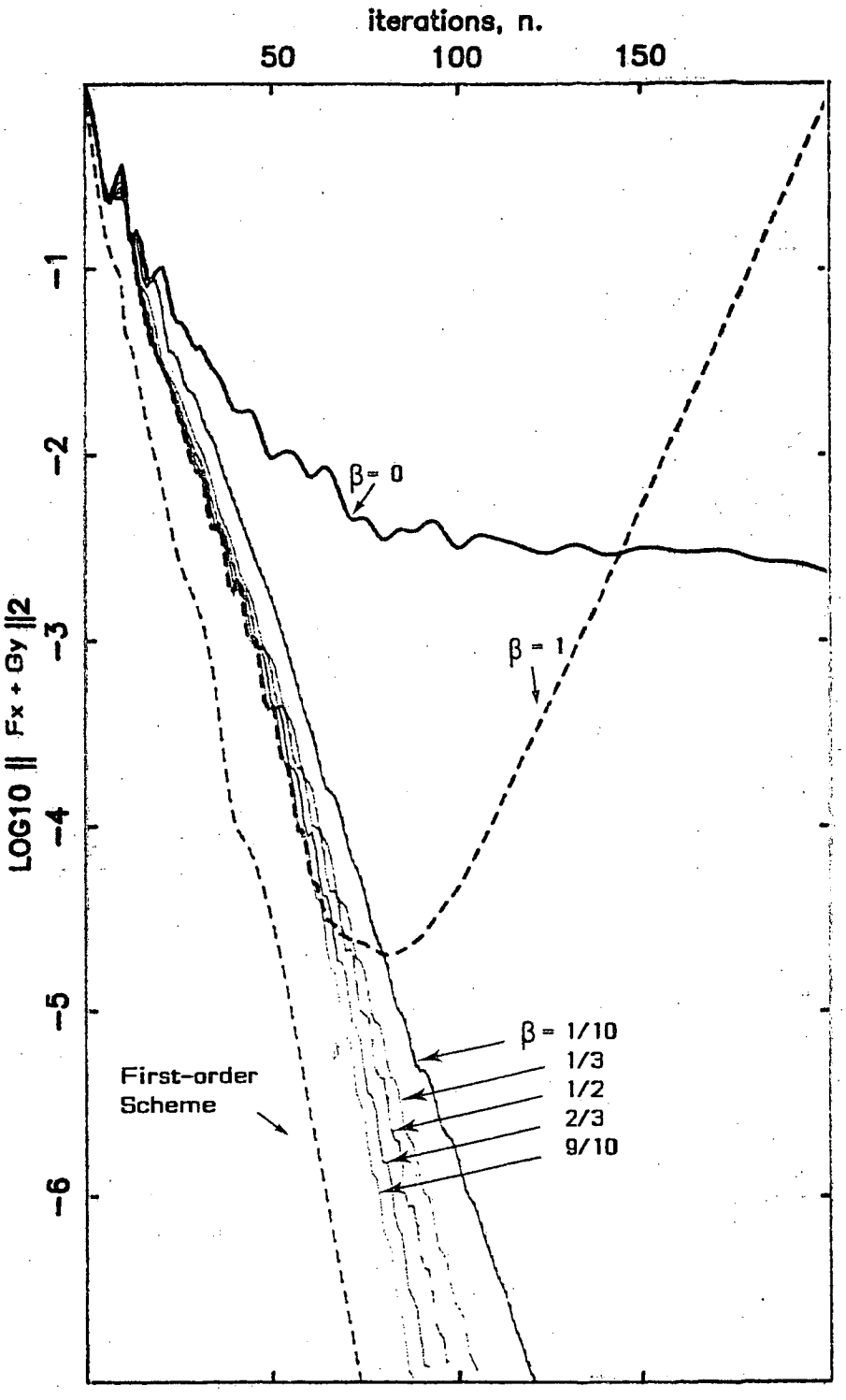

Figure 19. Convergence history for different values of the upwinding parameter β.

a - Subsonic Flow Regime.

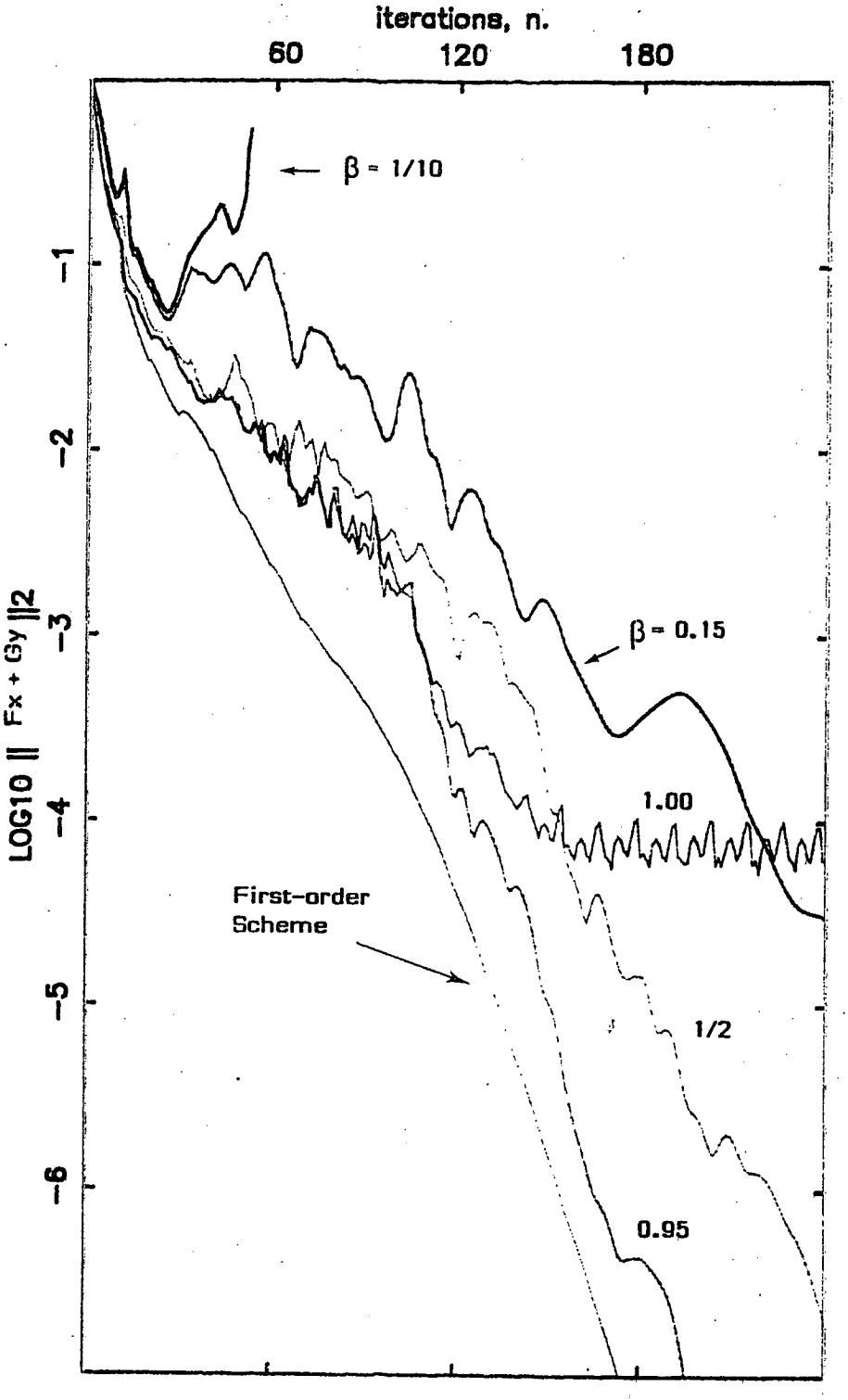

Figure 19. Continued

b - Transonic Flow Regime.

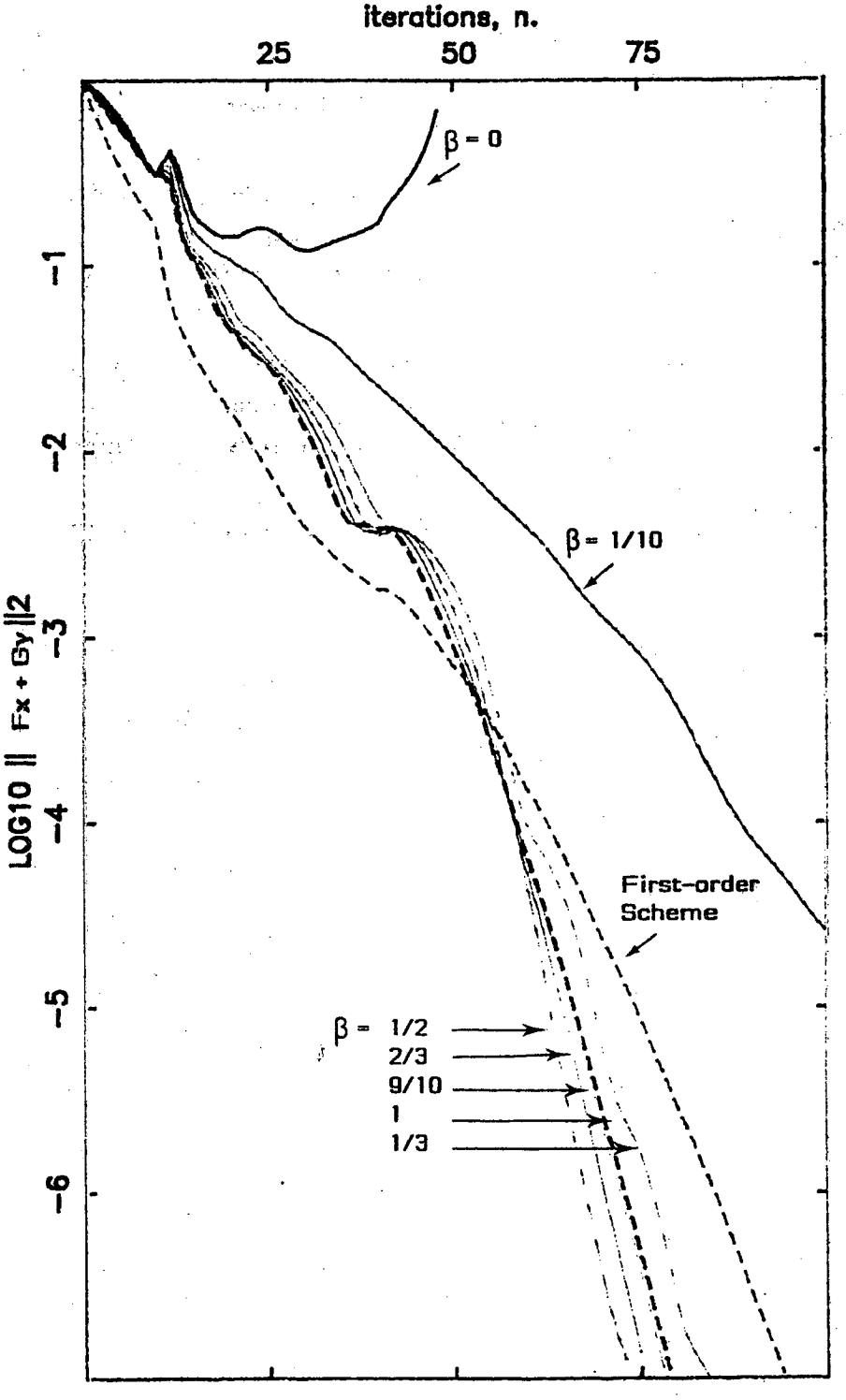

Figure 19. Continued

c - Supersonic Flow Regime.

5. CONCLUSIONS

In this report, the effects on iterative convergence of a defective amplification matrix, has been analyzed in a general context.

Practical results related to a simple linear problem often used to model hyperbolic equations, have been proved and then illustrated by numerical experiments. This revealed that utilizing a fully upwind differencing approximation in a certain implicit scheme could result in a paphological convergence.

Numerical experiments have been conducted with a finite-element program simulating a two-dimensional flow governed by the Euler equations. There also, difficulties have been observed to be associated with the usage of the fully upwind scheme.

Therefore to devise a fast second-order upwind scheme, we recommend to use one of the following two constructions:

(1) explicit phase: first-order approximation,

implicit phase: half-upwind second-order approximation,

(2) explicit and implicit phases: fully-upwind second-order approximations.

The second alternative constitute the most attractive challenge.

6. REFERENCES

[1] R.F. Warming, R.M. Beam and B.J. Hyett, "Diagonalization and Simultaneous Symmetrization of the Gas-Dynamic Matrices", Mathematics of Computation, Vol. 29, 132 (1975), pp.1037-1045.

[2] J.L. Steger, "Coefficient Matrices for Implicit Finite Difference Solution of the Inviscid Fluid Conservation Law Equations", Computer Methods in Applied Mechanics and Engineering 13 (1978), pp. 175-188.

[3] J.L. Steger, "Implicit Finite Difference Simulation of Flow about Arbitrary Geometries with Application to Airfoils", AIAA Paper 77-663 (1977).

[4] R. Peyret and T.D. Taylor, "Computational Methods for Fluid Flow", Springer, New York, 1983.

[5] D.A. Anderson, J.C. Tannehill and R.H. Pletcher, "Computational Fluid Mechanics And Heat Transfer", Hemisphere Publishing Corporation, McGraw-Hill Book Company, 1984.

[6] J.L. Steger and R.F. Warming, "Flux Vector Splitting of the Inviscid Gas-dynamic Equations with Applications to Finite-Diffference Methods", Journal of Computational Physics 40, 263-293 (1981).

[7] R. Beam and R.F. Warming, "An Implicit Finite-Difference Algorithm for Hyperbolic Systems in Conservation-Law-Form", Journal of Computational Physics, Vol. 22, Sept. 1976, pp. 87-110.

[8] J.M. Ortega and W.C. Rheinboldt, "Iterative Solution of Nonlinear Equations in Several Variables", Academic Press, New-York, 1970.

[9] J.L. Thomas, B. van Leer and R.W. Walters, "Implicit Flux-Split Schemes For The Euler Equations", AIAA Paper 85-1680, 1985.

[10] R.W. MacCormack, "Current Status of Numerical Solutions of the Navier-Stokes Equations", AIAA Paper 85-0032, 1985.

[11] G. Strang, "Linear Algebra and its Applications", Second Edition, Academic

Press, New York, 1980.

[12] R.S. Varga, "Matrix Iterative Analysis", Prentice-Hall, Inc., Englewood Cliffs, N.J., 1962.

[13] F. Fezoui and B. Stoufflet, "A Class of Implicit Upwind Schemes for Euler Simulations with Unstructured Meshes", INRIA Report, to appear.

7. APPENDIX A: Characteristic polynomial in the central-differencing case.

In this appendix, we calculate the characteristic polynomial of the matrix C_∞^c of (3.18), that is:

$$P_N^c(\lambda) = \det\,(C_\infty^c - \lambda I_N) \tag{a1}$$

in which I_N is the $N\times N$ identity matrix. Letting,

$$\mu \equiv 1 - 2\,\lambda \tag{a2}$$

results in:

$$C_\infty^c - \lambda I_N = \tfrac{1}{2}\begin{vmatrix} \mu+1 & -1 & & & & \\ 1 & \mu & -1 & & & \\ 1 & 0 & \mu & -1 & & \\ \vdots & \vdots & & \ddots & \ddots & \\ 1 & 0 & \cdots & 0 & \mu & -1 \\ 1 & 0 & \cdots & 0 & 1 & \mu-2 \end{vmatrix}_N \tag{a3}$$

in which the subscript N indicates that the matrix is $N\times N$. Expanding the above determinant along its first row yields:

$$2^N\,P_N^c(\lambda) = (\mu+1)\,f_{N-1}(\mu) + g_{N-1}(\mu) \tag{a4}$$

assuming the following definitions are made:

$$f_K(\mu) \equiv \det\begin{vmatrix} \mu & -1 & & & \\ 0 & \mu & -1 & & \\ \vdots & & \ddots & \ddots & \\ 0 & \cdots & 0 & \mu & -1 \\ 0 & \cdots & 0 & 1 & \mu-2 \end{vmatrix}_K \qquad (K \geq 2) \tag{a5}$$

and:

$$g_K(\mu) \equiv \det\begin{vmatrix} 1 & -1 & & & & \\ 1 & \mu & -1 & & & \\ 1 & 0 & \mu & -1 & & \\ \vdots & \vdots & & \ddots & \ddots & \\ 1 & 0 & \cdots & 0 & \mu & -1 \\ 1 & 0 & \cdots & 0 & 1 & \mu-2 \end{vmatrix}_K \qquad (K \geq 3) \tag{a6}$$

Developing $f_K(\mu)$ along the first column gives:

$$\begin{aligned} f_K(\mu) &= \mu f_{K-1}(\mu) \\ &= \mu^2 f_{K-2}(\mu) \\ &= \cdots \\ &= \mu^{K-2} f_2(\mu) \\ &= \mu^{K-2}[\mu(\mu-2)+1] \\ &= \mu^{K-2}(\mu-1)^2 \end{aligned} \tag{a7}$$

Analogously, developing $g_K(\mu)$ along its first row gives:

$$g_K(\mu) = f_{K-1}(\mu) + g_{K-1}(\mu) \quad (K \geq 4) \tag{a8}$$

Writing the above equation for all the values of K from 4 to $N-1$ and summing up give:

$$\begin{aligned} g_{N-1}(\mu) &= \sum_{K=4}^{N-1} \mu^{K-3}(\mu-1)^2 + g_3(\mu) \\ &= \mu(\mu-1)^2 \sum_{p=0}^{N-5} \mu^p + \det \begin{bmatrix} 1 & -1 & 0 \\ 1 & \mu & -1 \\ 1 & 1 & \mu-2 \end{bmatrix} \\ &= \mu(\mu-1)^2 \frac{1-\mu^{N-4}}{1-\mu} + [\mu(\mu-2)+1] + [(\mu-2)+1] \\ &= \mu(\mu-1)(\mu^{N-4}-1) + \mu^2 - \mu \\ &= \mu^{N-3}(\mu-1) \end{aligned} \tag{a9}$$

Utilizing (a4), (a7) and (a9) one obtains:

$$\begin{aligned} 2^N P_N^c(\lambda) &= (\mu+1)\mu^{N-3}(\mu-1)^2 + \mu^{N-3}(\mu-1) \\ &= (\mu-1)\mu^{N-3}(\mu^2-1+1) \\ &= (\mu-1)\mu^{N-1} \end{aligned} \tag{a10}$$

Finally, combining this result with the definition of μ given in (a2) yields the polynomial written in (3.19). ▪

8. APPENDIX B: Analysis of a modified central-differencing scheme.

In this appendix, we analyze a modified central-differencing scheme, obtained by replacing the last row of δ^c of (3.4) by the last row of δ^u of (3.5). In this way the truncation error is uniformly second-order, including the last grid-point ($x_j = x_N$). It is proved that this modification results in no significant improvement of the iterative convergence.

Thus, here,

$$\delta^c = Trid(-\tfrac{1}{2},0,\tfrac{1}{2}) = \tfrac{1}{2}\begin{bmatrix} 0 & 1 & & & & & \\ -1 & 0 & 1 & & & & \\ & -1 & 0 & 1 & & & \\ & & \cdots & \cdots & \cdots & & \\ & & & & -1 & 0 & 1 \\ & & & & 1 & -4 & 3 \end{bmatrix} \qquad \text{(b1)}$$

$$G^c_\infty = I - \delta_1^{-1}\delta^c = \tfrac{1}{2}\begin{bmatrix} 2 & -1 & & & & & \\ 1 & 1 & -1 & & & & \\ 1 & 0 & 1 & -1 & & & \\ \vdots & \vdots & & \cdots & \cdots & & \\ 1 & 0 & \cdots & 0 & 1 & -1 & \\ 1 & 0 & \cdots & -1 & 3 & 2 & \end{bmatrix} \qquad \text{(b2)}$$

$$G^c_\infty - \lambda I_N = \tfrac{1}{2}\begin{bmatrix} \mu+1 & -1 & & & & & \\ 1 & \mu & -1 & & & & \\ 1 & 0 & \mu & -1 & & & \\ \vdots & \vdots & & \cdots & \cdots & & \\ 1 & 0 & \cdots & 0 & \mu & -1 & \\ 1 & 0 & \cdots & -1 & 3 & \mu-3 & \end{bmatrix}_N \qquad \text{(b3)}$$

Thus (a4) still holds provided the definitions of the functions $f_K(\mu)$ and $g_K(\mu)$ are modified as follows:

$$f_K(\mu) \equiv \det \begin{bmatrix} \mu & -1 & & & & & \\ 0 & \mu & -1 & & & & \\ \vdots & & \cdots & \cdots & & & \\ 0 & \cdots & 0 & \mu & -1 & & \\ 0 & \cdots & 0 & 0 & \mu & -1 & \\ 0 & \cdots & 0 & -1 & 3 & \mu-3 & \end{bmatrix}_K \qquad (K \geq 3) \qquad \text{(b4)}$$

and:

$$g_K(\mu) \equiv \det \begin{bmatrix} 1 & -1 & & & & & \\ 1 & \mu & -1 & & & & \\ 1 & 0 & \mu & -1 & & & \\ \vdots & \vdots & & \cdots & \cdots & & \\ 1 & 0 & \cdots & 0 & \mu & -1 & \\ 1 & 0 & \cdots & 0 & 0 & \mu & -1 \\ 1 & 0 & \cdots & 0 & -1 & 3 & \mu-3 \end{bmatrix}_K \qquad (K \geq 4) \qquad \text{(b5)}$$

Developing $f_K(\mu)$ along the first column gives:

$$\begin{aligned} f_K(\mu) &= \mu f_{K-1}(\mu) \\ &= \mu^2 f_{K-2}(\mu) \\ &= \cdots \end{aligned}$$

$$= \mu^{K-3} f_3(\mu)$$

$$= \mu^{K-3} \det \begin{bmatrix} \mu & -1 & 0 \\ 0 & \mu & -1 \\ -1 & 3 & \mu-3 \end{bmatrix}$$

$$= \mu^{K-3}\{\mu[\mu(\mu-3)+3]-1\}$$

$$= \mu^{K-3}(\mu^3-3\mu^2+3\mu-1)$$

$$= \mu^{K-3}(\mu-1)^3 \qquad \text{(b6)}$$

Now, (a8) still apllies but only for $K \geq 5$, and summing this equation from $K=5$ to $N-1$ gives:

$$g_{N-1}(\mu) = \sum_{K=5}^{N-1} \mu^{K-4}(\mu-1)^3 + g_4(\mu) \qquad \text{(b7)}$$

$$= \mu(\mu-1)^3 \frac{\mu^{N-5}-1}{\mu-1} + \det \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & \mu & -1 & 0 \\ 1 & 0 & \mu & -1 \\ 1 & -1 & 3 & \mu-3 \end{bmatrix} \qquad \text{(b8)}$$

$$= \mu(\mu-1)^2(\mu^{N-5}-1) + \det \begin{bmatrix} \mu & -1 & 0 \\ 0 & \mu & -1 \\ -1 & 3 & \mu-3 \end{bmatrix} + \det \begin{bmatrix} 1 & -1 & 0 \\ 1 & \mu & -1 \\ 1 & 3 & \mu-3 \end{bmatrix}$$

$$= \mu(\mu-1)^2(\mu^{N-5}-1) + (\mu-1)^3 + \mu(\mu-3)+3+\mu-3+1$$

$$= (\mu-1)^2\mu^{N-4} - (\mu^3-2\mu^2+\mu) + (\mu^3-3\mu^2+3\mu-1) + \mu^2-3\mu+\mu+1$$

$$= (\mu-1)^2\mu^{N-4} \qquad \text{(b9)}$$

Finally,

$$2^N P_N^E(\lambda) = (\mu+1)\mu^{N-4}(\mu-1)^3 + (\mu-1)^2\mu^{N-4}$$

$$= (\mu-1)^2\mu^{N-4}(\mu^2-1+1)$$

$$= (\mu-1)^2\mu^{N-2}$$

$$= 4\lambda^2(1-2\lambda)^{N-2} \qquad \text{(b10)}$$

It is apparent from (b10) that:

Besides the eigenvalue $\lambda=0$ *which is double, the modified* G^c_∞ *matrix has only one eigenvalue,* $\lambda=\frac{1}{2}$ *of multiplicity* $N-2$. *(The spectral radius is thus* $\rho^c=\frac{1}{2}$.*)*

Let us now determine the eigenvectors:

- The vector $U=\{U_j\}$ is an eigenvector associated with the eigenvalue $\lambda=0$ if and only if:

$$0 = \frac{2U_1-U_2}{2} \tag{b11a}$$

$$= \frac{U_1+U_{j-1}-U_j}{2} \quad (\text{for } j = 3,4,\ldots,N) \tag{b11b}$$

$$= \frac{U_1-U_{N-2}+3U_{N-1}-2U_N}{2} \tag{b11c}$$

and this gives again the vector V of (3.11) only, one eigenvector is yet missing. Similarly, - The vector $U=\{U_j\}$ is an eigenvector associated with the eigenvalue $\lambda=\frac{1}{2}$ if and only if:

$$\frac{U_1-U_2}{2} = \cdots = \frac{U_1-U_j}{2} = \cdots = \frac{U_1-U_N}{2} = 0 \tag{b12a}$$

and

$$\frac{U_1-U_{N-2}+3U_{N-1}-3U_N}{2} = 0 \tag{b12b}$$

This gives U_j = a constant. Thus we find only one eigenvector associated with the multiple eigenvalue $\lambda=\frac{1}{2}$, that is again $U = V$ of (3.11). Therefore:

The modified G^c_∞ *matrix is defective,* $N-2$ *eigenvectors are missing, the largest Jordan block being of order N-3.*

Since N is large, we can combine the last two results with those of the Section 2 and conclude:

For the modified central differencing scheme ($\beta=0$), *and general initial guess, the iteration*

$$U^{n+1} = G^c_\infty\, U^n + b \tag{b13}$$

in which U^n *is the n-th iterate (an N-vector), b is a given N-vector, and N is considered large, is non-dissipative over a number of iterations of the order of* $2N$, *and then enters the final phase of convergence which is asymptotically like* $n^{N-3}/2^n$.

In conclusion, we observe that the slight modification of the differencing scheme at the boundary only brought a minor improvement on the asymptotic convergence rate.

9. APPENDIX C: A property of invariance of the condition number $\kappa(X)$.

In this appendix, we consider a general square matrix A (A corresponds to G_∞ in the main text), assumed to be reduced to diagonal form in two different ways, say,

$$A = X\,\Lambda\,X^{-1} = X'\,\Lambda'\,X'^{-1} \tag{c1}$$

in which Λ and Λ' are diagonal matrices containing the eigenvalues of the matrix A possibly differently ordered, and the matrices X and X' contain eigenvectors ordered correspondingly, possibly differently scaled but all normalized to 1. It is shown that the 2-norm condition numbers of the matrices X and X' are then identical.

For this, recall that the 2-norm of any square matrix T is equal to the square root of the largest eigenvalue of the matrix T^*T, where the superscript * indicates the adjoint (or transpose-conjugate). Consequently, the condition numbers in question are given by

$$\kappa(X) = \sqrt{\frac{\lambda_{max}(X^*X)}{\lambda_{min}(X^*X)}}\,,\quad \kappa(X') = \sqrt{\frac{\lambda_{max}(X'^*X')}{\lambda_{min}(X'^*X')}} \tag{c2}$$

Then we examine the relationships between the matrices Λ and Λ' and the matrices X and X'.

Let

$$\forall i \in \{1,2,\ldots,N\}\quad \lambda_i \equiv \Lambda_{i,i},\quad \lambda'_i \equiv \Lambda'_{i,i} \tag{c3}$$

and introduce the permutation p of $\{1,2,\ldots,N\}$ into itself such that

$$\forall i \in \{1,2,\ldots,N\},\quad \lambda'_i = \lambda_{p(i)} \tag{c4}$$

Then define the following permutation matrix:

$$P \equiv \{\, P_{j,k} \,\} \tag{c5}$$

where

$$\forall j,k \in \{1,2,\ldots,N\},\quad P_{j,k} \equiv \delta_{j,p(k)} \tag{c6}$$

δ being here the Kronecker symbol.

Let Δ be an arbitrary diagonal matrix, and j, k be two indices. We have:

$$(P^T\,\Delta\,P)_{j,k} = \sum_{m=1}^{N}\sum_{p=1}^{N} P^T_{j,m}\,\Delta_{m,p}\,P_{p,k}$$

$$= \sum_{m=1}^{N} P^T_{j,m}\,\Delta_{m,m}\,P_{m,k} \quad \text{(since } \Delta \text{ is diagonal)}$$

$$= \sum_{m=1}^{N} P_{m,j}\, P_{m,k}\, \Delta_{m,m}$$

$$= \sum_{m=1}^{N} \delta_{m,p(j)}\, \delta_{m,p(k)}\, \Delta_{m,m} \qquad \text{(c7)}$$

The general term in the above sum is equal to 0 anytime the index m differs from either $p(j)$ or $p(k)$. Therefore, if $j \neq k$, then $p(j) \neq p(k)$ (since p is injective), m cannot be equal to both, and none of these terms is nonzero; consequently, $(P^T \Delta P)_{j,k}=0$. If now, $j=k$, the above sum reduces to the term corresponding to $m=p(j)$, that is $\Delta_{p(j),p(j)}$. Thus, in general,

$$(P^T\, \Delta\, P)_{j,k} = \delta_{j,k}\, \Delta_{p(j),p(j)} \qquad \text{(c8)}$$

Using this with $\Delta=I$ yields

$$P^T P = I \qquad \text{(c9)}$$

showing that P^T is the inverse of P (-the column vectors of a permutation matrix are the vectors of the canonical basis ordered in a special way, and these form an orthogonal basis; it is therefore no surprise that the matrix P is found orthogonal-). Secondly, letting $\Delta=\Lambda$, yiels

$$P^T\, \Lambda\, P = \Lambda' \qquad \text{(c10)}$$

indicating that the matrix Λ' is obtained by permutation of the rows and the columns of the matrix Λ.

Let us now examine the relationship between the eigenvector matrices X and X'. The matrix X' is obtained from the matrix X by permutation of the columns only of the matrix X, corresponding to eigenvectors differently ordered, and by a different scaling of them, still respecting the constraint that they should all be of 2-norm equal to 1. This gives:

$$X' = X\, P\, S \qquad \text{(c11)}$$

in which S is a unitary diagonal matrix, that is:

$$S = Diag\,(\, e^{i\alpha_j}\,) \qquad \text{(c12)}$$

in which the α_j's are some real numbers and $i^2=-1$. This gives us:

$$S^{-1} = Diag\,(\, e^{-i\alpha_j}\,) = S^* \qquad \text{(c13)}$$

and

$$X'^*\, X' = S^*\, P^*\, X^*\, X\, P\, S$$

$$= S^{-1} P^{-1} X^* X P S$$

$$= (P S)^{-1} X^* X (P S) \qquad \text{(c14)}$$

where the fact that both matrices P and S are unitary has been used. It is directly apparent from (c14) that the matrices X^*X and X'^*X' are similar,

$$X'^* X' \sim X^* X \qquad \text{(c15)}$$

and thus have the same eigenvalues. Consequently, the condition numbers in (c2) are identical:

$$\kappa(X) = \kappa(X') \quad \blacksquare \qquad \text{(c16)}$$

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique